

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Philosophy

2013

Free Will and Abilities

Gadd, Nathan

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp015t34sj703>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

FREE WILL AND ABILITIES

NATHAN WILLIAM GADD

A THESIS

**PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY**

**RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
PHILOSOPHY**

Adviser: Gideon Rosen

November 2013

© Copyright by Nathan William Gadd, 2013. All rights reserved.

Abstract

My central aim in chapters 1 and 2 is to evaluate the Consequence Argument for incompatibilism about free will and determinism. In chapter 1, I present a modified version of the “modal argument” for incompatibilism and I develop a novel account of the argument’s central, quasi-technical idiom *being able to render a proposition false*. I begin chapter 2 by considering two alleged counterexamples to rule *BETA*, the Consequence Argument’s central rule of inference. I argue that these counterexamples fail, if the incompatibilist operates with a particular *luck dependent* notion of ability. Chapter 2 concludes with a discussion of David Lewis’ influential reply to the Consequence Argument. I argue, in particular, that Michael Huemer’s provocative objection to Lewis can be successfully deflected.

In chapter 3, I take a closer look at what the compatibilism debate is really about and I examine the luck dependent notion of ability in more detail. I begin by focusing on Kadri Vivhelin’s conception of the free will problem and I argue that it fails to isolate a unique problem. I conclude by sketching an alternative conception of the free will problem, one that sees the compatibilism dispute as a distinctively normative problem, a problem concerning the conditions under which an agent deserves to be harmed.

In chapter 4, with the normative conception of the free will problem in view, I examine the dispositional theory of ability, a theory that has recently been used to defend compatibilism. After explaining the motivation for the dispositional account, I argue that Michael Fara’s version of the theory fails to identify a form of ability that is necessary for moral responsibility. I then consider Michael Smith’s alternative version of the dispositional theory. While Smith’s view is promising, I argue that it faces serious challenges.

Table of Contents

Abstract	iii
Acknowledgements	v
Introduction	1
Chapter 1: The Consequence Argument	3
Chapter 2: Evaluating the Consequence Argument	28
Chapter 3: Two Conceptions of the Free Will Problem and Luck Dependent Abilities	58
Chapter 4: The Dispositional Theory of Ability	79
Bibliography	104

Acknowledgements

I am indebted to the following people for helpful discussion and correspondence concerning the issues in this thesis: Mark Balaguer, Joseph K. Campbell, Carl Ginet, Thomas Kelly, Barry Maguire, David Manley, Michael McKenna, Joe Rachiele, Whitney Schwab, Jack Spencer, Kadri Vihvelin, Gary Watson, Jack Woods, Gideon Yaffe and Luke Yovetich. I would like to add a special word of thanks to Michael Smith for several illuminating discussions over the years and for helpful feedback on an earlier draft. By far my greatest debt is to Gideon Rosen, who refined my thinking every step of the way.

Introduction

In chapter 1, I present a modified version of the so-called “modal argument” for incompatibilism. This modified argument, the Revised Consequence Argument (RCA), promises to establish a stronger form of incompatibilism than alternative versions. My main task in the chapter is to develop an account of the argument’s central, quasi-technical idiom *being able to render a proposition false*. I consider accounts of this idiom offered by Van Inwagen (1983) and David Lewis (1981), arguing that their accounts depart from the intuitive notion that the idiom is introduced to capture, namely the notion of an agent’s having control over a fact. I offer an alternative account of the idiom that does capture the intuitive notion in question. I conclude by briefly examining RCA’s premises and I argue that the compatibilism dispute reduces to a dispute over rule *BETA*, RCA’s central rule of inference.

I begin chapter 2 by answering Joseph Campbell’s (2008) objection to arguments like RCA. I then consider objections to rule *BETA*. An argument from McKay and Johnson (1996) is widely taken to demonstrate the invalidity of *BETA*. I argue that if the RCA defender operates with a particular *luck dependent* notion of ability, then the argument fails. I then consider an alleged counterexample to *BETA* from Kadri Vihvelin (1988), a counterexample that explicitly targets the luck dependent notion. I argue that Vihvelin’s counterexample fails and I show how the same response provided to Vihvelin’s case applies to an alleged counterexample from David Widerker (1987). After dispensing with these arguments, arguments that aim to demonstrate the invalidity of *BETA* on the incompatibilist’s own terms, I turn to the standard Lewisian reply to the consequence argument, a reply that presupposes compatibilism. I argue that, while the Lewisian reply is dialectically ineffective, it successfully deflects the incompatibilist challenge. I consider Michael Huemer’s ((2000) & (2004)) objection to the Lewisian reply and I explain how

the Lewisian ought to respond. I conclude by providing an answer, in the Lewisian spirit, to an argument offered by Van Inwagen (1983) in favor of *BETA*.

In chapter 3, I take a closer look at what the free will problem is really about and I examine the luck dependent notion of ability in more detail. I begin by focusing on Kadri Vihvelin's ((2008) & (2009)) conception of the compatibilism dispute and I argue that this conception fails to isolate a unique problem. After sketching a framework for analyzing luck dependent abilities, I argue that the traditional incompatibilist implementation of this framework renders *BETA* invalid and I provide an alternative that avoids this problem. I then argue that Vihvelin's conception of the compatibilism dispute, when understood as a dispute over luck dependent abilities, reduces it to a semantic question. I conclude by sketching an alternative conception of the free will problem, one that sees the compatibilism dispute as a distinctively normative problem, a problem concerning the conditions under which an agent deserves to be harmed.

In chapter 4, with the normative conception of the free will problem in view, I examine the dispositional theory of ability, a theory that has recently been used to defend compatibilism. After explaining the motivation for the dispositional account, I argue that Michael Fara's (2009) version of the theory fails to identify a form of ability that is necessary for moral responsibility. I then consider Michael Smith's (2003) alternative version of the dispositional theory. While Smith's view is promising, I argue that it faces serious challenges.

Chapter 1: The Consequence Argument

1. Introduction

Incompatibilists, philosophers who think that determinism is incompatible with free will, have traditionally defended their view by means of the following informal argument, *the remote past informal consequence argument* (RP):

RP: “If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us.”¹

Though there is surely something compelling about RP, the argument is not the hallmark of clarity and rigor. What rule or rules of inference, exactly, license the move from the premise that the remote past and the laws are not up to us to the conclusion that, if determinism is true, the consequences of these things are not up to us either? Why, exactly, should we accept that the laws of nature and events in the remote past are not *up to us*? And how should we interpret the argument’s central idiom of something’s being *up to us* in the first place?

Arguably the greatest advance made on the free will problem by analytic philosophers in the late twentieth century consisted in the transformation of the informal consequence argument into an argument that did not have the defects of RP. These new formal arguments inspired by RP were intended to make the form of reasoning implicit in the informal argument as perspicuous as possible, to provide a rigorous account of the *up to us* idiom, and to explain why we should accept the premises. As a result of this work, we now undeniably have a better understanding of what RP presupposes and of how its premises and inference rules might be defended. We are, thereby, better situated to examine potential compatibilist strategies of resistance. The ultimate hope is that, at the end of the day, when the master incompatibilist argument is in, even if we are

¹ Van Inwagen (1983), page 56.

unable to decisively prove that incompatibilism is true or false, we will have identified the premises and rules of inference upon which the incompatibilist relies. It can certainly seem as though we are close to having this master argument, one that crystallizes the guiding idea behind RP and lays bare the fundamental assumptions involved in it. We may be close to this ideal, tantalizingly so, but there is more work to be done.

* * * * *

One important, but relatively underexplored issue concerns the scope of incompatibilism. Traditional incompatibilists have not merely endorsed *the weak thesis*, the thesis that if determinism is actually true, then nobody actually has free will, but also *the strong thesis* that, necessarily, if determinism is true, then nobody has free will. Does the best version of the consequence argument purport to establish the strong thesis, the weak thesis, or some other thesis altogether?

Notice that RP explicitly relies on a contingent assumption, namely the assumption that there is a remote past, the assumption that for any agent there is some time in the remote past before that agent began to exist. This thesis is contingent for two reasons. First, it is possible that there be an *eternal agent*, an agent who has always existed at a world with no beginning. Second, it is possible that there be an *initial agent*, an agent who exists at the first moment of time. An argument that relies on the remote past assumption cannot establish that an eternal agent or an initial agent does not have free will. At best, such an argument can show that determinism is incompatible with free will *in possible worlds in which every agent has a remote past*. Can we dispense with the assumption that there is a remote past for all agents or is the assumption essential to the argument?²

² All of the points in this paragraph were first explicitly made, to my knowledge, by Joseph Campbell (2007) and (2008). He argues, mistakenly in my view, that attempts to dispense with the remote past assumption must fail. The

One might think that these issues concerning the scope of incompatibilism are idle, that they have no bearing on the issues that initially motivated our interest in the consequence argument, the issues of pressing human concern. We care about the consequence argument, the idea goes, because we care about whether we, actual human beings, have free will. And if we can establish the weak thesis, then we will have established that, if determinism is true, we lack free will. Why, then, bother at all with the issue of whether agents in other, non-actual deterministic worlds do not have free will?

First, it is not true that our interest in the consequence argument derives solely from an interest in the *actual freedom question*, the question of whether ordinary human beings are actually free, if determinism is true. If the oracle pronounced its verdict on the actual freedom question, we would then know whether or not we have free will, but many of us would remain unsatisfied. The reason is that we want to know *what it takes* to have free will and *why* determinism undermines free will if and when it does. The consequence argument is naturally understood as encoding the incompatibilist's fundamental insight, his explanation of why agents in deterministic worlds lack free will. It is not enough to have a true principle that entails that agents lack free will; we seek a principle that carves incompatibilism at its joints.

Second, RP is not the only informal argument that an incompatibilist might think to offer. In fact, there is an alternative informal argument that, *prima facie*, is as compelling an argument as RP, an alternative that dispenses with the remote past assumption (and so appears to establish a stronger form of incompatibilism than does RP). The availability of the alternative informal argument puts rational pressure on the incompatibilist who defends a version of RP: she must

RCA, presented in section 2, demonstrates that the assumption is not required. Campbell (2008) objects to an argument similar to RCA and I answer this objection in section 2 of chapter 2.

either say that the alternative informal argument really is as compelling as RP, as it appears to be, or she must identify its defects and show that they are not shared by RP.

Here is an alternative to RP, the *revised informal argument* (RI):

RI: If determinism is true, then our acts are the consequences of the laws of nature and the state of the world at any arbitrary time. In particular, relative to a time t , our acts are the consequences of the laws of nature and the state of the world at any time prior to t . But, for any arbitrary time t , it is not up to us at t what the state of the world is at any time prior to t , and it is not up to us at t what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us at t . And so for any time t' such that there is a time earlier than it, our acts are not up to us at t' .

On the face of it, RI is just as compelling as RP. The explicit premises of RI are as plausible as the premises of RP since events in the very recent past are no more up to us, after those events have occurred, than are events in the remote past. RP relies on an implicit “transfer” assumption, the principle, roughly, that if an event is the consequence of some factors that are not, will not be, and never have been up to us, then that event is not, will not be, and never has been up to us. RI relies on a different, but related transfer assumption, the claim that if an event is the consequence of some factors that are not up to us at some time t , then that event is not up to us at t . So long as this time-indexed principle is as plausible as the non-indexed version, as it appears to be, RI is as secure an argument as RP.

Like RP, RI does not entail that initial agents cannot have free will. But, if RI is sound, it does show that eternal agents in deterministic worlds lack free will. This is why RI is an improvement over RP. It constitutes a step towards identifying the incompatibilist’s fundamental explanatory principle, the principle that explains why (according to the incompatibilist) determinism undermines free will when it does.

My aim in this chapter is to make progress at identifying the master version of the consequence argument. I begin, in section 2, by presenting the *Revised Consequence Argument* (RCA), an argument inspired by RI that is modeled on Van Inwagen's so-called "modal version" of the consequence argument based on RP. In section 3, I provide commentary to RCA. A crucial feature of RCA is that it dispenses with the ordinary, but vague, notion of something's being *up to us* utilized by RP and RI, replacing it with the quasi-technical idiom *being able to do something that would render a proposition false*. Many contemporary writers on the consequence argument do not provide explicit accounts of this technical idiom at all, presumably taking for granted, mistakenly, that its meaning is transparent. And discussions that do attempt to elucidate this technical idiom are both incomplete and importantly confused. I proceed by first discussing three existing accounts of the idiom. I then provide an alternative account that better captures the intuitive notion that the technical idiom is introduced to capture. In section 4, I provide a brief discussion of RCA's central premises and I argue that the dispute over RCA reduces to a dispute over the validity of the argument's central rule of inference, rule *BETA*.

2. *Revised Consequence Argument (RCA)*³

Let ' $\Box$ ' stand for the metaphysical necessity operator. Let '<' stand for the *earlier than* relation. Let 'G' express an arbitrary true proposition. Let 'L' express the *law stating proposition*, the proposition conjoining all and only the laws of nature. Let t and t' be variables ranging over times. Relative to an assignment of some time t^* to ' t ' in ' $h(t)$ ', ' $h(t)$ ' is a true proposition describing the entire intrinsic state of the world at t^* . Let p be a variable ranging over formulas. Let ' $N(p, t)$ ' be defined as ' p and nobody is able, at t , to do something that

³ Brueckner (2008) offers a similar argument in response to Campbell (2007).

would render p false'.⁴ Where P and Q are arbitrary sentences, here are three inference rules governing 'N':

ALPHA: From $\Box P$ infer $\forall t N(P, t)$.

BETA: $\forall t$ From $N(P, t)$ and $N((P \rightarrow Q), t)$ infer $N(Q, t)$.

AGG: $\forall t$ From $N(P, t)$ and $N(Q, t)$ infer $N((P \wedge Q), t)$.⁵

Where t_1 is an arbitrary time later than t_0 , here is the Revised Consequence Argument (RCA):

- | | |
|--|-------------------------------|
| (1) $t_0 < t_1$ | [Stipulation] |
| (2) $\Box \forall t ((h(t) \wedge L) \rightarrow G)$ | [Determinism] |
| (3) $\forall t' \forall t (((h(t) \wedge L) \rightarrow G), t')$ | [2, <i>ALPHA</i>] |
| (4) $\forall t \forall t' (t < t' \rightarrow N(h(t), t'))$ | [Relative Fixity of the Past] |
| (5) $N(h(t_0), t_1)$ | [1, 4] |
| (6) $\forall t N(L, t)$ | [Fixity of the Laws] |
| (7) $N(L, t_1)$ | [1, 6] |
| (8) $N((h(t_0) \wedge L), t_1)$ | [5, 7, <i>AGG</i>] |
| (9) $\forall t N(((h(t_0) \wedge L) \rightarrow G), t)$ | [1, 3] |
| (10) $N(((h(t_0) \wedge L) \rightarrow G), t_1)$ | [1, 9] |
| (11) $N(G, t_1)$ | [8, 10, <i>BETA</i>] |

The argument is a conditional proof. Given that $t_0 < t_1$ and given that determinism is true, we derive the result that nobody is able, at t_1 , to do something that would render G false. Since G is any arbitrary, true proposition and since the choice of t_1 was arbitrary except for the condition that there must be a time earlier than t_1 , the argument, if sound, establishes that for all times t such that there is a time earlier than t , nobody is able, at t , to do something that would render any true proposition false.

⁴ Nearly every writer on the consequence argument has been convinced that rule *BETA* is invalid, if the N -operator is defined in terms of the 'would'-counterfactual. I take up this issue in chapter 2.

⁵ *AGG* must be a valid inference rule if *ALPHA* and *BETA* are. Suppose that $N(P, t_1)$ and $N(Q, t_1)$ but $\sim N((P \wedge Q), t_1)$. By *ALPHA*, since $\Box (P \rightarrow (Q \rightarrow (P \wedge Q)))$, we infer $\forall t N((P \rightarrow (Q \rightarrow (P \wedge Q))), t)$. Thus, $N((P \rightarrow (Q \rightarrow (P \wedge Q))), t_1)$. By *BETA* and $N(P, t_1)$ we infer $N(Q \rightarrow (P \wedge Q)), t_1)$. By *BETA* and $N(Q, t_1)$ we then infer $N((P \wedge Q), t_1)$. Contradiction.

ALPHA is uncontroversial. If *BETA* is a valid inference rule, then the argument is valid. The only premises available for dispute are (4) and (6).

3. The N-operator

3.1 Preliminary remarks

The central premises of RCA, and its conclusion, are stated in terms of the N-operator, an operator defined in terms of a quasi-technical idiom, the notion of an agent's being able to do something that would render a proposition false. We need to introduce a technical idiom of this sort because, while free will is typically understood in terms of an agent's having the ability to perform certain actions, determinism is a thesis relating propositions. Thus, to connect our ordinary free will claims with the thesis of determinism, we need some way of describing our ordinary claims about an agent's abilities in terms of that agent's bearing a certain relation to propositions. On one common usage of 'free will', an agent S has free will, at t, with respect to some action A only if S is able, at t, to do otherwise than A. The following principle states the intended connection between this ordinary concept of free will and our technical idiom:

Connecting Principle (CP): S is able, at t, to do otherwise than she does only if S is able, at t, to do something that would render some true proposition false.

If CP is true, then RCA, if sound, shows that no agent in a deterministic world is able, at any non-initial time t, to do otherwise than she does.

The meaning of the quasi-technical idiom, *being able, at a time, to do something that would render a proposition false*, though not perfectly transparent, is at least somewhat intuitive.

Following Van Inwagen (1983), let us introduce the notion with examples. Consider the following propositions: that $2 + 2 = 4$, that the earth rotates, that Joan of Arc was burned at the stake, and that my parents met in 1974. These are facts that are not, have not been, and never will be under my control. I do not have a choice about them. In terms of the technical idiom, we

say that I am, always have been, and always will be unable to do something that would render these true propositions false.

By contrast, consider the following propositions: that I did not visit Moscow in 2001, that my arm is not up, and that my bedroom window will not break tonight. These are facts that are, at one time or another, under my control. I was able, in 2001, to purchase a ticket and fly to Moscow. So I was able, in 2001, to do something that would render false the proposition that I have never been to Moscow. I am now able to stop typing and raise my arm. So I am now able to do something that would render it false that my arm is not up. I will be able, when I am staring at my bedroom window tonight, rock in hand, to break the window. So I will be able to do something that would render it false that the window is not broken.⁶

We can speak more generally of an event's rendering a proposition false at a world. The event of the rock's crashing through the window, at w , renders false at w the proposition that the window is not broken. Jones' formation of the belief, at w , that nothing has ever happened renders it false at w that nothing has ever happened. These events do not render false, at any world, the falsehoods that Joan of Arc was not burned at the stake and that $2 + 2 = 5$.

In light of the foregoing examples, I believe that the notion of an event's rendering a proposition false at a world and the derivative notion of an agent's being able to do something that would render a proposition false are tolerably clear. But it would be better if we had a more precise account. In the rest of this section, I first consider three extant accounts of these notions and I argue that these accounts fail to capture the intuitive notion that I have just introduced. I then offer a novel definition that does capture the intuitive data.

⁶ Of course, if RCA is sound, then I do not really have these abilities, if the actual world is deterministic.

3.2 The Weak Sense and the Strong Sense

I begin by employing a distinction made by Lewis (1981). The claim that an agent is able, at t , to do something that would render a proposition false can be understood in a weak sense (WK) and a strong sense (ST):

WK: An agent has the weak ability, at t , to do something that would render a proposition P false if and only if the agent is able, at t , to do something such that, had he done it, an event e would have occurred and rendered P false.

ST: An agent has the strong ability, at t , to do something that would render a proposition P false if and only if the agent is able, at t , to do something such that, had he done it, an event e would have occurred such that (i) e is either the agent's action or an effect of his action and (ii) e renders P false.

Note that WK and ST are not intended to be analyses of the ability to perform an action. Rather, they are accounts of the technical idiom that make use of the ordinary notion of an agent's being able to perform an action. Since sentences of the form 'S is able to X' are context sensitive, expressing different propositions on different occasions of use, being clear about what weak and strong abilities are requires that we specify an intended interpretation of sentences of the form 'S is able to X'. But specifying such an intended interpretation raises a number of thorny issues that are orthogonal to the central purpose of this chapter. I am going to put off discussion of this issue until later, but the reader should bear in mind that, while I am using the language of ability throughout this chapter, it should not be clear precisely how I am using it.

WK and ST are stated in terms of the 'would'-counterfactual. Taking the Lewisian semantics as our model, WK and ST are equivalent to WEAK and STRONG:

WEAK: An agent has the weak ability, at t , to do something that would render a proposition P false if and only if the agent is able, at t , to do something such that all of the nearest worlds w in which he does it are worlds in which an event e (i) occurs and (ii) renders P false at w .

STRONG: An agent has the strong ability, at t , to do something that would render a proposition P false if and only if the agent is able, at t , to do something X such

that all of the nearest worlds w in which he does it are worlds in which an event e (i) occurs, (ii) is the agent's Xing or an effect of his Xing, and (iii) renders P false at w .

On the Lewisian semantics, which worlds are the “nearest worlds” in the evaluation of a counterfactual is, in general, a context sensitive matter. But I stipulate here that we understand all of the counterfactuals in this paper in terms of the so-called “standard resolution”. In general, a counterfactual of the form *if e had occurred at t , it would have been that C* is true at a world w just in case C is true at all of the nearest worlds relative to w where e occurred at t . On the standard resolution, if w is deterministic, the nearest worlds relative to w where e occurs at t are worlds that, if they differ from w at all, share w 's history up until the occurrence of an event prior to t that violates a law of nature at w . The occurrence of such an event at a nearby world w^* is a miracle at w^* relative to w . After the miracle event occurs at w^* , everything occurs in conformity with the laws of w . I will also assume *strong centering*: if the antecedent of a counterfactual is true at w , then the nearest antecedent world relative to w is w and no other world.⁷

We generate different versions of WEAK and STRONG by providing different definitions of *event e renders p false at w* . In the following section, I will consider three definitions of the idiom. Every candidate definition will be assigned a number n . WEAK_n and STRONG_n are the versions of WEAK and STRONG that result from replacing *e renders p false at w* with Definition n . Possession of the weak_n and strong_n ability, at a time, to do something that would render a proposition P false is equivalent to possessing the ability on the right hand side of the bi-conditional in WEAK_n and STRONG_n , respectively.

⁷ For details of the Lewisian semantics see Lewis (1979).

3.3. Three Definitions of *e renders p false at w*

3.3.1 Definition 1

Here is the modal definition employed by Lewis (1981):

Definition 1: *e renders P false at w* if and only if *e* occurs at *w* and necessarily (if *e* occurs, then *P* is false)

Definition 1 has intuitively correct results in a wide range of cases. Consider the true proposition, proposition *O*, that I did not just raise my arm. In normal cases, cases in which I am not paralyzed or handcuffed or sleeping and so on, we typically think that I was able just now to raise my arm. Suppose that this is right. Then I was able just now to perform an action, the arm-raising event, and each of the nearest worlds *w* in which I did so are worlds where the arm-raising event renders *O* false at *w* in the sense of Definition 1. That is, every such *w* is such that the arm-raising event occurs at *w* and necessarily (if the arm-raising event occurs, *O* is false). Thus, I have the weak₁ and strong₁ abilities to do something that would render *O* false.

Van Inwagen (1983) argued that Definition 1 is unacceptable on the basis of the following case:

NASSER: All of Nostradamus' predictions were correct, including his prediction that the Sphinx would not be destroyed. Nasser was able, at *t*₀, to destroy the Sphinx. Nostradamus made all of his predictions before *t*₀.

Intuitively, Nasser has free will, at *t*₀, with respect to the fact that all of Nostradamus' predictions were correct. After all, Nasser had the ability, at *t*₀, to destroy the Sphinx and Nostradamus predicted before *t*₀ that the Sphinx would not be destroyed. We thus want the result that Nasser had the ability, at *t*₀, to render it false that all of Nostradamus' predictions were correct. But, the objection continues, WEAK₁ and STRONG₁ do not have this consequence. Note that, while Nasser was able to destroy the Sphinx, it is not true that necessarily (if the destruction event occurs, it is false that all of Nostradamus' predictions are correct).

Nostradamus might have made only one prediction, that the Sphinx would be destroyed, at a world in which the destruction event occurred. Or Nostradamus might have existed at a world in which the destruction event occurs, but made no predictions at all. Thus, the NASSER objection goes, the account does not capture a necessary condition for being able to do otherwise. So CP is false, where that principle is understood in terms of either weak₁ or strong₁ abilities.

Though Van Inwagen and others following him hold that NASSER demonstrates the unacceptability of Definition 1⁸, I believe that the argument just given moves too quickly. It is natural to defend the account in the following way. Suppose that Nasser had destroyed the Sphinx. There are many ways that we might choose to describe what Nasser did. He destroyed a statue, he upset the Egyptians, he hammered away at a giant sculpture, and he did the thing reported on page 421 of the history book on my desk. Moreover, the description ‘Nasser saw to it that Nostradamus’ predictions were not all correct’ is a perfectly legitimate description of what Nasser did. Let ‘the i-event’ be an abbreviation of this description. Nasser was able, at t_0 , to do something – namely cause the i-event – such that all of the nearest worlds w in which he caused it are worlds such that necessarily (if the i-event occurs, then it is false that all of Nostradamus’ predictions are correct). Thus, Nasser has the weak₁ and strong₁ abilities, at t_0 , to do something that would it render false that all of Nostradamus’ predictions are correct. Call this reply to the NASSER objection the *Description Strategy*.

I said that the original argument against Definition 1 on the basis of NASSER moves too quickly. From the fact that possibly (if the destruction event occurs, then all of Nostradamus’ predictions were correct), what follows? It is not obvious that we can conclude that NASSER is not able, at t_0 , to do something that would render it false that all of Nostradamus’ predictions

⁸ Schnieder (2004), for example, rejects Definition 1 on the basis of a parallel case.

were correct. What the Description Strategy suggests is that if we are careful about selecting a description of Nasser's act, if we choose 'the i-event' instead of 'the destruction event', then Definition 1 does in fact generate the correct verdict in NASSER.

The Description Strategy raises a problem, however, the *too many falsehoods problem*. The reply requires that we take a liberal view about what kinds of descriptions adequately describe an agent's actions. In particular, it requires that we accept certain gerrymandered descriptions, like 'Nasser's seeing to it that Nostradamus' predictions were not all correct', as referring to Nasser's destruction of the Sphinx. The problem is not that we have to consider these descriptions to be admissible – they *are* legitimate descriptions of Nasser's action. But once we accept this liberal view concerning which descriptions are admissible, there is the following problem.⁹ Suppose that Nasser destroyed the Sphinx at w . Let Q be any false proposition at w . Consider the following description of what Nasser did: 'Nasser destroyed the Sphinx while not- Q '. Given the liberal view, this should be an admissible description of Nasser's action. But then it appears that we also have the consequence that Nasser is weakly₁ and strongly₁ able, at t_0 , to do something that would render false *every* false proposition. For all of the nearest worlds in which Nasser destroys the Sphinx while not- Q are worlds such that necessarily (if Nasser destroys the Sphinx while not- Q , then Q is false). More generally, if an agent does anything that he is able to do, then he is weakly₁ and strongly₁ able to render false every false proposition.¹⁰

One response to the too many falsehoods problem, the response Lewis would have given, denies that that Definition 1, correctly interpreted, has this problematic consequence. The Lewisian reply begins by insisting that Definition 1 is to be understood as making a *de re* modal claim about events. To see this, consider the nearby world w in which Nasser destroyed the

⁹ Thanks to Gideon Rosen and Jack Woods for discussion.

¹⁰ Note that incompatibilists, like Van Inwagen (1983), typically do not deny that an agent is able to perform actions at deterministic worlds; they merely deny that an agent is able to do otherwise at such worlds.

Sphinx. Let Q be the contingent falsehood at w that Nate is wearing a blue shirt right now. We can agree that ‘destroying the Sphinx while not- Q ’ describes something that Nasser did at w . But, on the de re interpretation of Definition 1, the claim that this event renders Q false is to be understood as follows: Nasser’s act of destroying the Sphinx while not- Q , act e , is such that necessarily (if e occurs, then Q is false). To examine this necessity claim, we must have an account of e ’s essential properties. And it is plausible to hold that *being such that not- Q* is merely an accidental feature of e , a feature that Nasser’s act could have lacked. To put it another way, it is plausible to hold that there is no act that Nasser performed which is such that it could not have occurred if Nate were not wearing a blue shirt right now. If this is correct, then even though ‘destroying the Sphinx while not- Q ’ refers to Nasser’s act, it is not the case that *that* event is impossible with Q .

The de re interpretation of Definition 1 blocks the problematic result that, if an agent is able to do anything, he is weakly₁ and strongly₁ able to render false every falsehood. Of course, on the de re interpretation, the Description Strategy fails to answer the NASSER objection. The Description Strategy appealed to a gerrymandered description of Nasser’s act, the description ‘seeing to it that Nostradamus’ predictions were not all correct’. But ‘the act of ensuring that Nostradamus’ predictions were not all correct’, like ‘the act of destroying the Sphinx while not- Q ’, does not plausibly describe what Nasser did by way of an essential property of that act. That is, Nasser’s act of seeing to it that Nostradamus’ predictions were not correct, act e , could have occurred at a world in which Nostradamus’ only prediction was that I would be typing right now. Thus, it is not the case that necessarily (if e occurred, then it is false that all of Nostradamus’ predictions were correct).

Here is a second problem for the de re interpretations of Definition 1: it entails that any event e that occurs at w renders U false, where U is any necessary falsehood. Thus, if an agent S has the ability to do anything at all, S has the weak₁ and strong₁ ability to do something that would render U false. This is counterintuitive. The mere fact that I am able to raise my arm surely does not show that I have control over the fact that it is false that $2 + 2 = 5$. In the intuitive sense, then, I am unable to render it false that $2 + 2 = 5$.

Let me summarize the argument of this section so far. The NASSER objection is designed to show that Definition 1 is unacceptable: we want the result that Nasser is (weakly and strongly) able, at t_0 , to render it false that all of Nostradamus' predictions are correct, but, allegedly, Definition 1 does not have this consequence. But the argument given for this conclusion moves too quickly. The Description Strategy provides a natural reply; it appears to show that Definition 1 really does generate the correct verdict in NASSER. The Description Strategy ultimately fails, however, because it generates the *too many falsehoods* problem. To avoid this problem in a Lewisian spirit, we ought to interpret Definition 1 as making a de re claim about events. But then the Description Strategy fails, so long as there is no event that is essentially an event of seeing to it that all of Nostradamus' predictions were correct. Moreover, Definition 1 entails that, if an agent does anything that he is able to do, then he is weakly₁ and strongly₁ able to do something that would render false every necessary falsehood. Thus, the de re interpretation of Definition 1 fails to capture the intuitive notion of being (weakly or strongly) able to render a proposition false.

I move next to consider a second modal definition.

3.3.2 Definition 2

Definition 2 was proposed by Van Inwagen (1983) in response to the original NASSER objection. Van Inwagen's thought about the NASSER case is essentially that, in determining whether Nasser's destroying the Sphinx would have falsified the proposition that all of Nostradamus' predictions were correct, we have to somehow keep fixed Nostradamus' actual predictions. Van Inwagen thus attempts to deal with the case by "building the past" into the account of an event rendering a proposition false. Here is his alternative definition:

Definition 2: e renders P false at w if and only if e occurs at w and necessarily (if e occurs and the past is identical to the past at w up until the occurrence of e , then P is false)

Nasser is weakly₂ and strongly₂ able to render it false that all of Nostradamus' predictions were correct. For Nasser is able to do something – destroy the Sphinx, event d – such that all of the nearest worlds w in which d occurs are worlds where (i) d is caused by his act and (ii) is such that necessarily (if d occurs and the past is identical to the past at w up until the occurrence of d , it is false that all of Nostradamus' predictions are correct). This is so because all of the nearest worlds in which Nasser destroys the Sphinx are worlds in which Nostradamus predicted that the Sphinx would be destroyed before that event occurred.

There are three problems with Definition 2. First, like Definition 1, Definition 2 predicts that I am weakly₂ and strongly₂ able to render it false that $2 + 2 = 5$ by raising my arm. Second, as Van Inwagen recognizes, it entails that by raising my arm, I am weakly₂ and strongly₂ able to render it false that Joan of Arc was not burned at the stake. For if I raise my arm, then I am able to raise my arm. Thus, I am able to do something – raise my arm – such that had I done it, the arm-raising event e would have occurred where e is my action and is such that necessarily (if the past is the actual past up until e occurs, then it is false that Joan of Arc was not burned at the stake).

Third, there is reasoning analogous to that given on the basis of NASSER against Definition 1 for the conclusion that Definition 2 must also be rejected. Consider the case NASSER+:

NASSER+: Nasser was able, at t_0 , to destroy the Sphinx, but did not. Max will make the incorrect prediction at some time later than t_0 that the Sphinx will never be destroyed. Nasser knows, at t_0 , that Max, at some time later than t_0 , will make this prediction.

It is plausible to suppose that Nasser has free will, at t_0 , with respect to the proposition that all of Max's predictions will be correct because he has the ability, at t_0 , to destroy the Sphinx and he knows that Max will predict that it will not be destroyed. But suppose that Max could have avoided making this prediction. Then it is not true that Nasser is able, at t_0 , to do something – destroy the Sphinx, event d – such that all of the nearest worlds w in which he causes d is a world such that necessarily (if d occurs and the past is as it is at w up until the occurrence of d , then it is false that all of Max's predictions are correct). Definition 2 does not have a decisive advantage over Definition 1. If the NASSER objection undermines Definition 1, NASSER+ undermines Definition 2.

3.3.3: Definition 3

Schnieder (2004) reports that Lewis suggested the following account to Terence Horgan:

WOULD: An agent S has the ability, at t , to do something that would render P false if and only if S is able, at t , to do something such that, had he done it, P would have been false.

WOULD is equivalent to $WEAK_3$, where Definition 3 is the following:

Definition 3: e renders P false at w if and only if P is false at w .

WOULD nicely handles the NASSER objection. The act of destroying the Sphinx is such that had it occurred, it would have been false that all of Nostradamus' were correct. This is so because the nearest worlds in which Nasser destroys the Sphinx are all worlds in which Nostradamus predicted what he actually predicted. But though this account does handle the

original NASSER objection, it falls prey to the too many falsehoods problem. Suppose that I just raised my arm and was thus able to do so. The nearest world in which I raised my arm is, by the strong centering assumption, the actual world since I actually raised my arm. Thus, it is true that, had I raised my arm, Q would have been false, where Q is any actual falsehood. Thus, I am weakly₃ able to do something that might render Q false, for any false proposition Q.

3.4. An Alternative Proposal

I have considered three definitions of the idiom *e renders P false at w* and three corresponding accounts of the derivative idiom *agent S is able, at t, to do something that would render P false*. I have argued that each definition constitutes a radical departure from the intuitive notion that we hope to capture. I here begin to develop a different proposal that does capture the intuitive notion.

Consider the following pairs of claims:

(1a) Joe's belief that grass is purple is false because grass is green.

(1b) Grass is green because Joe's belief that grass is purple is false.

(2a) The proposition that Nate did not break the window is false because Nate broke the window.

(2b) Nate broke the window because the proposition that Nate did not break the window is false.

(1a) and (2a) are true, while (1b) and (2b) are false. What explains why Joe's belief is false is that the world is not the way that Joe's belief represents it as being. In particular, Joe's belief represents the world as being such that grass is purple and this representation is false, or incorrect, because grass is green. Grass' being green, I hold, explains why Joe's belief is false. And so (1a) is true. For (1b) to be true it would have to be the case that the fact that Joe's belief

is false explains why grass is green. But grass' being green does not depend on representations of the world in this way and so it does not depend on Joe's representing the world as he does. So (1b) is false.

The explanation for the truth of (2a) and the falsity of (2b) is essentially the same as the explanation just given for the truth of (1a) and the falsity of (1b). This is because, like beliefs, propositions are representations of the world, whether you think that propositions essentially and intrinsically represent the world by their very nature or whether you think that propositions represent the world because they are interpreted by us as doing so.¹¹ Propositions are true, when they are, because the world is the way the proposition represents it as being. And they are false, when they are, because the way they represent the world as being is not the way that the world is. In particular, the truth of (2a) is explained as follows: Nate broke the window and that fact explains why the proposition that Nate did not break the window, a representation of the world as being such that Nate did not break the window, has the property of being false. For (2b) to be true it would have to be the case that the fact that Nate broke the window is explained by the fact that a representation of Nate's breaking the window has a certain property. But Nate's breaking the window does not depend on representations of it in this way. So (2b) is false.¹²

The world's being the way that it is explains why certain representations of the world are true and why others are false. This platitude can be used to provide an alternative account of the notion of *e's rendering p false at w* that captures the intuitive sense of the notion:

Definition 4: *e* renders *P* false at *w* if and only if *P* is false at *w* because *e* occurred.

Return to the original examples I used to introduce the technical idiom. Since I was able, in 2001, to fly to Moscow, I was weakly₄ and strongly₄ able to do something in 2001 that would

¹¹ See Soames (2008) and Merricks (2008) for further discussion.

¹² This way of introducing the 'because' notion and the explanation offered of the intuitive contrast between (1a)/(1b) and (2a)/(2b) follows Soames (2008).

render false the proposition that I have never been to Moscow. Since I was just able to raise my arm I was just weakly₄ and strongly₄ able to do something that would render it false that my arm is not up. I had the ability to do something that would render false the proposition that I have never been to Moscow because we think that I was able to fly there and that, had I done so, my flying to Moscow would have *explained why* it is false that I have never been to Moscow. It would have been false that I have never been to Moscow, had I flown there, *because* I flew there. Had I just raised my arm, that I just raised my arm would have been the *reason* that it is false that my arm was not raised. It would have been false that I did not raise my arm, had I raised it, *because* I raised it. The alternative proposal thus captures the cases that we used to introduce the notion.

This account blocks the consequence that Nasser, by being able to destroy the Sphinx, is thereby weakly₄ and strongly₄ able to do something that would render it false that Joan of Arc was not burned at the stake. It blocks this consequence because the nearest worlds in which Nasser destroys the Sphinx, or does anything else that he is able to do, are not worlds in which that proposition is false because of Nasser's action or any event caused by that action. Moreover, the account entails that Nasser is weakly₄ and strongly₄ able, at t_0 , to do something that would render it false that all of Nostradamus' predictions are correct. Nasser's destroying the Sphinx would explain why that proposition is false.

The NASSER case shows that the intended sense of explanation must be that of *partial* explanation. While it is perhaps plausible to think that the event of the window's breaking completely explains the falsity of the proposition that the window is not broken, the destruction event does not completely explain the falsity of the proposition that all of Nostradamus' predictions are correct. The occurrence of the destruction event explains the falsity of that

proposition, but a complete explanation must make reference to the fact that Nostradamus predicted, in particular, that the Sphinx would not be destroyed.

The account also captures the fact that nobody has control over the fact that it is false that $2 + 2 = 5$. Nothing anyone is able to do might cause an event such that $2 + 2 = 5$ is false because that event occurred. Thus, nobody is weakly₄ or strongly₄ able, at any time, to do something that would render it false that $2 + 2 = 5$.

The central drawback to Definition 4 is that the notion of explanation expressed by ‘because’ is a bit fuzzy. I introduced the notion by (i) exploiting it to explain the intuitive contrast between (1a)/(2a) and (1b)/(2b) and (ii) describing certain paradigm cases in which the occurrence of an event explains the falsehood of a certain proposition. And I believe that my use of the notion has made it sufficiently intelligible. But I will not offer a view about the nature of this explanatory notion. Some theorists, like Kit Fine and Gideon Rosen, hold that the notion of explanation in question is a relation of ontological dependence or of metaphysical priority.¹³ But many philosophers suspicious of esoteric metaphysical pictures of this kind are inclined to deny that there is any metaphysical relation expressed by ‘because’, the so-called grounding relation, and instead prefer to give our explanatory talk some sort of conceptual, non-metaphysical spin.¹⁴ While I am sympathetic to the “esoteric” metaphysical picture, I do not think that the RCA defender is committed to this controversial metaphysics merely in virtue of the fact that she has given an account of the N-operator in terms of definition 4.

I trust that the explanatory notion expressed by ‘because’ has been made sufficiently clear. From now on, I will drop the subscript notation and will presuppose the understanding of the N-operator along the lines of WEAK and STRONG, with *e renders P false* defined by Definition 4.

¹³ For details see Rosen (2009) and Fine (2001)

¹⁴ See, for example, Hofweber (2009).

Section 4: The Fixity of the Past Premise

4.1 Relative Fixity of the Past

At this point, the reader should think that we have settled on two versions of RCA, one version with the N-operator interpreted along the lines of WEAK and another where it is interpreted along the lines of STRONG. In this section, I offer what I take to be the strongest defense of premise 4, the Relative Fixity of the Past premise. If the RCA defender gives this defense of premise 4, then she has good reason to use STRONG instead of WEAK. One consequence of this defense of the Relative Fixity of the Past premise is that RCA does not establish the strongest form of incompatibilism: the defense I offer for this premise does not apply to certain deterministic worlds and RCA can thus establish, at best, a restricted form of incompatibilism.

Premise 4 of RCA, the Relative Fixity of the Past premise, is plausible because we do not have the powers of backwards causation. That is, we do not have the power, at a time t , to cause the occurrence of an event at any time t' prior to t , whether t' is some time in the remote past relative to t or just an instant before t . Intuitively, the reason that nobody is now able to render it false that Joan of Arc was burned at the stake in 1431 or that nobody now has power over the proposition that my arm was not raised one second ago is that nobody now has the ability to cause an event that explains the falsity of those propositions. If the fact that we lack backwards causal powers is what explains why we accept that we do not, now, have free will with respect to these kinds of propositions, then we should defend the Relative Fixity of the Past premise by appealing to this feature of our abilities.

However, it is unclear how we would do this if we defined the N-operator in terms of WEAK. An agent's possession of the weak ability to do something that would render a proposition P

false implies that she is able to do something such that her doing it requires that an event that falsifies P has occurred. But it does not require that the agent is able to perform or cause this event. STRONG, by contrast, does require this. Thus, on the basis of STRONG, we can justify the Relative Fixity of the Past premise, as we intuitively do, by noting that, at any time t, nobody is able, at t, to cause an event at any time t' earlier than t that would be an h (t')-breaking event. Importantly, I believe that the typical compatibilist will accept the Relative Fixity of the Past premise, for the reasons just given, when the N-operator is interpreted along the lines of STRONG.

I claimed in section 3 that RCA, if sound, establishes that no agent at a deterministic world is able, at any non-initial time t, to do otherwise than she does. But given the defense of premise 4 offered here, RCA can only establish the following weaker conclusion: no agent in a deterministic world *that does not permit backwards causation* is able, at any non-initial time t, to do otherwise than she does.

4.2 Fixity of the Laws

It is worth noting, first, that Lewis originally introduced the “strong” sense of the *able to render false* idiom in order to capture a sense in which, even by the compatibilist's lights, nobody is able to render false the law-stating proposition. The compatibilist will, indeed, accept the Fixity of the Laws premise given my account of the N-operator in terms of STRONG.

In the rest of this section, I will mention three assumptions about laws of nature that are required for the Fixity of the Laws premise to be plausible. I will not examine these assumptions in detail, but it is worth noting them because they are far from uncontroversial.

First, the premise requires that there actually are laws of nature.¹⁵

¹⁵ For skepticism about this see Van Fraassen (1989).

Second, it appears to require that no Humean view of the laws is true. A minimal requirement for an account of the laws to be Humean is that the account maintains that the laws of nature are laws because of what actually occurs and, therefore, that what actually occurs does not occur because of what the laws are. Suppose it is a law that particles with the same charge repel one another. On the Humean view, when particle p1 and particle p2 repel, where p1 and p2 are both negatively charged, this is not explained by the fact that it is a law that particles with the same charge repel one another. Rather, that the law is a law is explained, in part, by the fact that p1 and p2 behave as they do. It does not seem that, on an account of laws of this type, the laws constrain what an agent is able to do. The Relative Fixity of the Laws premise, however, thus appears justifiable only given a conception of laws according to which they are facts that are settled independently of our behavior. If it is a law, at our world w, that nothing travels faster than light (law L1), then, on the non-Humean view, I cannot cause any event that might constitute an L1-breaking event at w because L1 is settled independently of my behavior. I am not certain that there is no way to defend the Fixity of the Laws premise, given a Humean account of laws, but I do not see how to do so.¹⁶

Third, for RCA to have any interesting implications it must be the case that the laws are exceptionless generalizations. Suppose, as some philosophers maintain, that there can be exceptions to deterministic laws. If we think that there can be exceptions to the laws, then we might offer an alternative definition of ‘determinism’. Say that a set of laws J is deterministic just in case any two worlds that are intrinsically alike at some moment of their history and that perfectly conform to J are intrinsically alike at every moment of their history. But premise (2) of RCA does not follow from this definition of determinism. It might be true that, if w1 and w2, worlds that are intrinsically alike at some time t, perfectly conformed to the laws of their world,

¹⁶ See Beebe & Mele (2002) for further discussion.

then they would be intrinsically alike at all times. It might nonetheless be the case that w_1 and w_2 have the same laws, but that while w_1 perfectly conforms to the laws, w_2 does not, i.e. some event occurs at w_2 that constitutes an exception to the laws of w_2 . Suppose, for example, that w_1 and w_2 are both governed by J , but that, at (w_1, t_1) a particle p moves right and at (w_2, t_1) p moves left. Then, where R is the proposition that p moves left, and $t_0 < t_1$, it is not true that $\Box ((h(t_0) \wedge J) \rightarrow R)$.

6. Conclusion

The aim of this chapter was to make progress at identifying the best version of the consequence argument. The argument presented here is an improvement over earlier versions of the argument because it dispenses with the assumption, not essential to the incompatibilist's guiding thought, that there is a remote past. The argument, if sound, shows that any non-initial agents existing at a deterministic world lack free will, so long as those worlds do not permit backwards causation. Most importantly, we now have a clear, intuitive account of the central technical idiom employed by the argument. This account allows us to offer a compelling justification for the Relative Fixity of the Past premise, one that the compatibilist ought to accept. And the Relative Fixity of the Laws premise appears plausible so long as the laws are exceptionless generalizations, settled independently of our behavior. Taking this picture of the laws of nature for granted, the dispute between the compatibilist and the incompatibilist over RCA reduces to a dispute over the validity of *BETA*, the argument's central rule of inference.

Chapter 2: Evaluating the Consequence Argument

1. Introduction

In section 2 of this chapter, I answer an objection to RCA from Joseph Campbell (2008). This will help to clarify the account of an agent's having the strong ability to do something that would render a proposition false. I then move to a discussion of rule *BETA*. There are two types of objections to *BETA*, objections that do not presuppose compatibilism and objections that do. I discuss objections of the former type in sections 3 – 5 and an objection of the latter type in section 6. In particular, in section 3, I consider a case that has convinced nearly every writer on the consequence argument that Van Inwagen's original formulation of the transfer principle is invalid, a formulation given in terms of the 'would'-counterfactual rather than the 'might'-counterfactual. I argue that, given a particular notion of *luck dependent abilities*, the counterexample fails. In section 4, I argue that an alleged counterexample to *BETA* from Kadri Vihvelin fails. In section 5, I argue that an alleged counterexample to *BETA* from David Widerker fails. In section 6, I consider the classic compatibilist treatment of *B* deriving from David Lewis (1981) and consider an objection to the Lewisian view from Michael Huemer ((2000) & (2004)). I argue that the Lewisian can effectively deflect Huemer's objection. I then employ the Lewisian strategy utilized to answer Huemer's objection to answer an argument offered by Van Inwagen (1983) in favor of *BETA*.

Section 2: Campbell's Objection

Campbell (2008) objects to arguments like RCA, arguments that appeal to the relative fixity of the past as opposed to the absolute fixity of the remote past, on the grounds that they do "not have anything to say about the past in particular."¹⁷ We can, he says, "just as easily talk about

¹⁷ Campbell (2008) page 266.

the *necessity of the relative future* as we can about the necessity of the relative past.”¹⁸ His idea, then, is that we can give an argument structurally similar to RCA that utilizes the relative fixity of the future premise – the premise that $\forall t \forall t' (t > t' \rightarrow N(h(t), t'))$ – instead of the relative fixity of the past premise. If Campbell is right, we should find an argument in terms of the relative fixity of the future just as compelling as RCA.

Campbell is wrong. Intuitively, the future is *not* relatively fixed, a fact that is captured by the account I have offered of the *ability to render false* idiom.¹⁹

Why does Campbell hold that the future is relatively fixed? Suppose that, at t_2 , Joe will wear shoes instead of sneakers. Let P2 be the true proposition that Joe will not wear sneakers at t_2 .

Campbell says:

“[P]resumably I am able now, at t_1 , to render it false that I will not wear sneakers at t_1 but that doesn’t mean that I have a choice now about whether I will not wear sneakers at t_2 ; it doesn’t mean that I have a choice now about whether [P2]. On the face of it, it seems that I am unable now to render it false that I will not wear sneakers at t_2 ...so I have no choice about whether [P2]. Thus, relative necessity applies to the future as well as to the past” (266).

Campbell is reasoning in the following way: he is able (or so he claims), at t_1 , to render it false that Joe will not wear sneakers at t_1 since he is able, at t_1 , to put them on at t_1 .²⁰ But he is unable, at t_1 , to render P2 false because he is unable, at t_1 , to cause an event to occur at t_1 that completely

¹⁸ Campbell (2008) page 266.

¹⁹ Note, however, that even if Campbell were right that, intuitively, the future is relatively fixed, the objection would not cast doubt on the *soundness* of RCA. Campbell himself appears to accept that both the past and future are relatively fixed and so the objection is not that RCA has a false premise. But Campbell also does not think that his objection shows that the argument relies on a faulty rule of inference. Nonetheless, though Campbell’s objection does not cast doubt on the soundness of RCA, it is still a cause for concern. Recall that we are aiming to provide an argument that encapsulates the heart of incompatibilist thinking. And it is thus an important question whether the relative fixity of the past premise can be replaced by a premise stating that the future is relatively fixed.

²⁰ This is what Campbell is thinking – he says, after all, “...I am able now, at t_1 , to render it false that I will not wear sneakers at t_1 ” – but I think that this is incorrect. If Campbell is not wearing sneakers at t_1 , then he is not able, at t_1 , to render it false that he is not wearing sneakers at t_1 . To render that proposition false he would have to cause the event of his wearing sneakers at t_1 . But, if his sneakers are not on at t_1 , he is not able, at t_1 , to cause that event since putting on sneakers is not an instantaneous action, i.e. it takes time to put on sneakers. At best, Campbell is able, at t_1 , to render it false that he is not wearing sneakers at some time soon after t_1 .

explains the falsity of P2. I agree that Campbell is unable to do this. But it does not follow that, intuitively, the future is relatively fixed.

Consider an analogous case. A basketball player is able, when he is standing at the foul line preparing to shoot a free throw at t , to sink the free throw. We do not think that the player is unable, at t , to sink the free throw just because he cannot, at t , cause the ball to go through the hoop *at t* . We think that the player is able, at t , to send the ball through the hoop at some time t' later than t . And this is why we say that the player is able, at t , to sink the free throw. Similarly, we should not hesitate to say that an agent has the ability, at a time t , to wear sneakers at some later time, just because he is not already wearing them at t . Given that he is able, at t , to choose to wear sneakers, he is able, at t , to do something that would cause him to be wearing sneakers at some time after t . In such cases, it is perfectly appropriate to say that the agent is able, at t , to be wearing sneakers at some time after t . Thus, we should not agree that we pre-theoretically think that Campbell was unable, at t_1 , to be wearing sneakers at t_2 . Intuitively, he has the ability, at t_1 , to do something that would render P2 false. And so, intuitively, the future is not relatively fixed.

My account of the *able to render false* idiom captures the fact that, intuitively, the future is not relatively fixed. Campbell is able, at t_1 , to put on sneakers at t_2 . And, had he done so, the event of his putting on sneakers at t_2 would have (i) occurred at t_2 , (ii) been his action, and (iii) explained the falsity of P2. Thus, Campbell has the weak and strong abilities, at t_1 , to render P2 false.

Section 3: COIN

Consider COIN, a variant of an example from McKay and Johnson (1996):

COIN: Raul is able, at t , to flip a fair coin – a coin with an objective chance of .5 of landing tails and with an objective chance of .5 of landing heads – but he never flips it, and nobody else has access to the coin.

Several writers believe that COIN constitutes a counterexample to RCA's rules of inference. Joseph Campbell, for example, claims that the COIN constitutes a "decisive" counterexample to the consequence argument's central rule of inference.²¹ Kadri Vihvelin claims that "agglomeration is uncontroversially invalid" and she uses COIN to justify this claim.²² Van Inwagen says: "the case imagined [COIN] is undeniably a counterexample to Agglomeration. Agglomeration is therefore invalid, and the invalidity of [BETA] follows from the invalidity of Agglomeration."²³

I will first present an argument, on the basis of COIN, for the conclusion that AGG is invalid. I will then present an argument, on the basis of COIN, for the conclusion that BETA is invalid.

In COIN, Raul is not (weakly or strongly) able, at t , to do something that would render it false that *the coin did not land heads* (H) – he is not able to ensure that the coin does not land heads. Raul is not (weakly or strongly) able, at t , to do something that would render it false that *the coin did not land tails* (F) – he is not able to ensure that the coin does not land tails. But he is (weakly and strongly) able to do something – flip the coin – that would render it false that *the coin did not land heads or tails* ($H \wedge F$). By flipping the coin, Raul ensures that the coin lands either heads or tails. Thus, AGG is invalid since $N(H, t)$ and $N(F, t)$, but $\sim N((H \wedge F), t)$. Since AGG must be valid, if ALPHA and BETA are valid, and since nearly everyone holds that ALPHA is valid, nearly everyone holds that the argument against the validity of AGG amounts to an indirect argument against the validity of BETA.

Here is a direct argument for the invalidity of BETA on the basis of COIN. Let H be the proposition that the coin does not land heads. Let K be the proposition that the coin is not flipped. The coin does not land heads and Raul is not (weakly or strongly) able, at t , to do

²¹ Campbell (forthcoming).

²² Vihvelin (2011).

²³ Van Inwagen (2000).

something that would render it false that the coin does not land heads ($N H, t$). While $H \rightarrow K$ is true, Raul is not (weakly or strongly) able, at t , to do something that would render it false that, if the coin does not land heads, then the coin is not flipped ($N (H \rightarrow K), t$). To (weakly or strongly) falsify the material conditional Raul must be able to do something that guarantees that the coin is flipped and that the coin does not land heads. But though Raul is able to flip the coin, it might land tails and it might land heads. So Raul is not able to ensure that the coin does not land heads and is flipped. So he is not (weakly or strongly) able to do something that would render $H \rightarrow K$ false. But he is (weakly and strongly) able, at t , to do something that would render K false since he is able to flip the coin. Thus, *BETA* is invalid since $N (H, t)$ and $N ((H \rightarrow K), t)$, but $\sim N (K), t$.

I will next argue that there is a sense of ‘able’, the luck dependent sense, according to which the above arguments fail.²⁴

Consider $N (H, t)$, a premise of both the indirect and direct arguments. The defense of this premise that I offered above proceeds by noting that, while Raul is able to flip the coin, it is not true that, had he flipped the coin, it would have landed heads. But note that it does not immediately follow that he is strongly or weakly unable, at t , to render H false. There might be some action X that he is able to perform, an action other than flipping the coin, which is such that had he done X , an event e would have (i) occurred, (ii) been his action or a causal consequence of his action, and (iii) rendered H false. In fact, it is natural to think that there is such an action available to Raul, namely the act of flipping heads. Of course, Raul does not have a special method of flipping the coin that will guarantee, or make it especially likely, that the coin lands heads. He does not have any special skills that would make it reasonable for him to

²⁴ Thanks to Gideon Rosen and Jack Spencer for helpful discussion.

bet that he would flip heads instead of tails if he tried. Still, he is able to flip heads and, had he done so, H would have been false because of this action.

Why think that Raul is able to flip heads? Note first that the MIGHT HAVE DONE IT principle entails that, in COIN, Raul is able to flip heads:

MIGHT HAVE DONE IT: S is able to Y if (i) he is able to X and (ii) had S Xed, he might have Yed.

It is stipulated that Raul is able to flip the coin and, had he flipped it, he might have flipped heads. Thus, by the MIGHT HAVE DONE IT principle, Raul was able to flip heads.

Here is a related argument for the conclusion that Raul was able to flip heads. Suppose that COIN is true of the actual world @. Consider world w, which differs from @ in that Raul flips heads by exercising the ability he possesses at @ to flip the coin. At w, Raul has the ability to flip heads since an agent is able to perform an action if he does perform that action. But the only difference between Raul at w and Raul at @, is that, at w, Raul exercised his ability to flip the coin, an ability that Raul had at @. But if this is the only difference between Raul at w and Raul at @, then they must have all of the same abilities. So, at @, Raul is able to flip heads.

You might reject the above argument by holding that, at w, Raul did not have the ability to flip heads, even though he did flip heads. That is, you might accept (1):

(1) Raul flipped heads but he was not able to flip heads.

The thought is perhaps that, in general, if it is just a matter of luck that an agent performs a particular action, then he is not able to perform that action. And Raul was lucky to flip heads at w. If this is right, then we should also reject the MIGHT HAVE DONE IT principle because, even if it is true that an agent might have done something Y by exercising his ability to X, his doing Y might have been lucky and would then not have been an exercise of the ability to Y.

There is something right about this objection. Raul is no magician. If he tries to pick the ace of spades from the deck, he will succeed roughly two percent of the time. If he tries to walk through the wall, he will almost certainly fail. And if he tries to flip heads, he will only succeed fifty percent of the time. If and when Raul does pick the ace of spades, walk through the wall, or flip heads, his successes are lucky. Compare Raul to Saul, a magician who possesses reliable methods for picking the ace of spades, walking through walls, and flipping heads. It is natural to characterize the difference between Raul and Saul by saying that Saul has, while Raul lacks, the ability to pick the ace of spades, walk through walls, and flip heads. In this sense of ‘able’, (1) is true.²⁵

But while there is a reading of (1) on which it is true, there is a clear reading of (1) according to which it is false. In this alternative sense of ‘able’, the fact that an agent does something amounts to a proof that he is able to do that thing, regardless of whether or not the agent’s doing that thing was lucky. It does not matter how unlikely it is that an agent performs the action and it is not required that an agent possess a reliable method for doing what he does: if an agent does something, then he was able to do it. Raul might have no special method for picking the ace of spades from the deck. But if he does pick the ace of spades, then he was able to do it. If he walks through the wall, then, no matter how small the probability was that he performs that action, he was able to walk through the wall. And if he flips heads, then he was able to do so.

The sense of ‘able’ according to which (1) is false – the sense in which Raul is able to flip heads, pick the ace of spades, and walk through walls, given that he does these things – picks out a *luck dependent ability*. Raul does not possess a special method for reliably doing these things

²⁵ I doubt that noting that the contrast between Raul and Saul is sufficient for fixing a determinate sense of ‘able’. There may be many senses in which Saul is, while Raul is not, able to flip heads, pick the ace of spades, and walk through walls. Select any such determinate sense. Then, in that sense, (1) is true of Raul at w.

and so, when he does them, his successes are lucky. Given the sense of ‘able’ on which it picks out a luck dependent ability, we should accept the MIGHT HAVE DONE IT principle.

If we use the sense of ‘able’ according to which Saul has, while Raul lacks, the ability to flip heads, then the arguments on the basis of COIN for the invalidity of *AGG* and *BETA* succeed.²⁶

However, if we interpret ‘able’ in the luck dependent sense, then COIN is unproblematic.

$N(H, t)$ is a premise of both arguments. But Raul has the luck dependent ability, at t , to flip heads and, had he done so, H would have been false because he flipped heads.

It is worth noting a crucial difference between the notion of luck dependent abilities I am employing here and Carl Ginet’s (1990) notion of “broad” abilities. While Ginet’s broad abilities are similar to my luck dependent abilities, his notion departs from mine in a crucial way. The difference is important because Ginet’s broad abilities cannot avoid a version of the counterexample on the basis of COIN. Consider the following passage from Ginet (1990):

“The distinction I have in mind depends on whether or not an agent’s needing a significant amount of luck in order to perform a certain action is taken to rule that out of those open to an agent. Suppose that I am a very inexpert dart-thrower, but I happen by sheer luck to throw the dart into the bulls-eye. Or suppose that I do not know the combination for a lock I wish to open, but I happen by sheer luck to dial the right combination on my third try. One can have conflicting intuitions

²⁶ Here is one well-known fix for the RCA proponent who operates with a more robust notion of ability, one that precludes luck. We modify WEAK and STRONG as follows:

WEAK*: An agent has the weak ability, at t , to do something that **might** render a proposition P false if and only if the agent is able, at t , to do something X such that one of the nearest worlds w in which he does it is a world in which an event e (i) occurs and (ii) renders P false at w .

STRONG*: An agent has the strong ability, at t , to do something that **might** render a proposition P false if and only if the agent is able, at t , to do something X such that one the nearest worlds w in which he does it is a world in which an event e (i) occurs, (ii) is the agent’s X ing or an effect of his X ing, and (iii) renders P false at w .

With the N -operator defined in terms of WEAK* and STRONG*, COIN is not a counterexample to *BETA*. Raul is able to flip the coin and one of the nearest worlds in which he does so is a world in which the coin lands heads as a consequence of action and renders it false that the coin does not land heads. Thus, given the account in terms of WEAK* and STRONG*, it is not the case that $N(H, t)$.

about such cases. On the one hand, it is natural to say that, since my action of hitting the bulls-eye with the dart (or of opening the lock) actually occurred, it must have been open to me to perform. Even in the far likelier case where I fail to hit the bulls-eye (fail to dial the combination that opens the lock), we might accept the statement that with enough luck I could have hit it (could have opened it). On the other hand, it also seems natural to say that, since I had no skill at throwing darts (had no idea what the combination of the lock was) and it was sheer luck that I hit the bulls-eye (dialed the right combination), it was not open to me to do it, not something I could do.”²⁷

These remarks are compatible with everything that I have said. However, Ginet’s implementation of the idea departs from mine in an important way. Ginet argues, in light of the intuitive distinction here, that we need to distinguish between what he calls “broad” (or luck-dependent) abilities and “narrow” abilities. The narrow ability to X requires that you possess some skill in Xing. So Raul lacks the narrow ability to flip heads. The broad ability to X is defined in terms of the narrow ability as follows:

“At t it was open to S (broad sense) to perform action A if and only if at t it was open to S (narrow) sense to perform some action such that, if S were to choose to perform that action, then S would perform action A (and it is not ruled out, nor is it required, that if S were to perform the action whereby S would perform A, S’s performing A would be by luck rather than through skill or knowledge).”²⁸

Ginet’s account avoids the counterexample on one version of COIN, but not on another. Here is the version of COIN that Ginet’s account can handle. Suppose that Raul has a way of flipping the coin – he is narrowly able to flip the coin that way - such that had he flipped it that way, the coin would have landed heads. Raul therefore has the broad ability to flip heads and the counterexample is blocked.

But suppose that, in COIN, the coin flip is chancy in the following way: no matter how the coin strikes the ground, it might have landed heads and it might have landed tails. So there is no particular way of flipping the coin such that, had Raul flipped it that way, it would have landed

²⁷ Ginet (1990), page 96.

²⁸ Ginet (1990), page 97.

heads. On Ginet's view, it follows that Raul is not able to flip heads. For there is no way of flipping the coin such that Raul had the narrow ability to flip it that way, i.e. did not require luck to flip it that way, and such that had he flipped it that way, it would have landed heads. By contrast, MIGHT HAVE DONE IT entails that Raul is able to flip heads in this case. This version of COIN is a counterexample to *BETA* given Ginet's "broad" ability: the argument against *BETA* stated above succeeds, given Ginet's notion. For Raul is not strongly able, in the broad sense, to render it false that the coin does not land heads, nor that (if it does not land heads, then it is not flipped). But he is strongly able, in the broad sense, to render it false that coin is not flipped. But this version of COIN is not a counterexample given my notion of a luck dependent ability since Raul is strongly able, in the luck dependent sense, to do something that would render it false that the coin does not land heads.²⁹

Section 4: BETTY

In this section I consider Kadri Vihvelin's (1988) case BETTY, a case that she presents as a problem for *BETA* given the notion of luck dependent abilities³⁰:

BETTY: Everyone in Betty's town has a number in an indeterministic lottery. For a person to win the lottery the person's number must be picked and the person must pay a fee. Betty is able, at t_0 , to pay the fee that is required to win the lottery, but she does not pay it. The picking of the winning number is causally isolated from Betty's behavior. Her number is not picked and she does not win the lottery.

Here is an argument, on the basis of BETTY, for the invalidity of *BETA* where N is interpreted along the lines of STRONG. Let V be the proposition that Betty's number isn't picked. Let W be the proposition that Betty doesn't win. The picking of the number is, by stipulation, causally isolated from Betty. So causing her number to be picked is not something

²⁹ Ginet pointed me to the above passages in personal communication.

³⁰ Vihvelin explicitly attributes the luck dependent notion to Ginet and, as I've argued above, his notion departs from mine. However, the success (or failure) of the BETTY counterexample does not turn on the difference between Ginet's broad ability and my luck dependent ability.

that Betty is able to do and nothing that she is able to do would cause her number to be picked. Thus, since V is true we have $N(V, t_0)$. To be able, at t_0 , to render $V \rightarrow W$ false in the strong sense, she must be able, at t_0 , to bring about an event that falsifies the conditional. But it is impossible to win if her number isn't picked. So she is not strongly able, at t_0 , to render the conditional false. Thus, since $V \rightarrow W$ is true we have $N(V \rightarrow W, t_0)$. However, it is natural to think that Betty is strongly able, at t_0 , to render W false. Notice that if we accept that *had Betty paid the fee, then she might have won*, then by MIGHT HAVE DONE IT she is able to win. She is able in the luck dependent sense to win the lottery and, had she done so, W would have been false because of her winning. Thus, *BETA* must be invalid since we have that $N(V, t_0)$ and $N(V \rightarrow W, t_0)$, but $\sim N(W, t_0)$.³¹

I am going to argue that, in the luck dependent sense, Betty is strongly unable, at t_0 , to do something that would render it false that she does not win the lottery. I will first explain Timothy O'Connor's argument for denying that Betty has the ability in question. I argue that O'Connor's argument fails. I then present my reasons for agreeing with his conclusion.

O'Connor (1993) argues that Betty is unable to win the lottery, in the luck dependent sense, in the following way. He asks us to first imagine that there is only one fee-payment location available to Betty and that "the proprietor [let's call him 'Steve'] of that establishment elected to close up shop and go fishing", so that the location is closed at t_0 . The location's being open at t_0 is thus a necessary condition on Betty's winning the lottery. We suppose that Steve had the ability to refrain from fishing and to open the store instead. O'Connor claims that we would not say that Steve "could have made it the case that Betty won the lottery, given only that the action

³¹ Note that if we operate with a more robust form of ability, a form of ability that precludes luck and requires that an agent possess a reliable method for success, then we should not think that Betty is able to render it false that she does not win.

available to him was a causally necessary factor and the putative fact that it might have been the case (if he had opened the store) that Betty decided to pay the fee and the draw turned up her number.” Since there is a necessary condition on Betty’s winning, namely that her number is picked, a condition Steve has no control over, O’Connor denies that Steve is able to make it the case that Betty wins the lottery. And he denies that Betty is able to win the lottery, for the same reason: there is a necessary condition on her winning, namely her number’s being picked, and Betty has no control over whether or not this condition obtains.

According to O’Connor an agent has the luck dependent ability to render a proposition P false when there is some sequence of bodily movements available to the agent sufficient for the falsity of P, where the agent may be “unaware of what that sequence is that will do the trick.”³² To see what O’Connor has in mind, consider Bill, who does not know the combination to the safe. Bill is able to turn the dial from 37 to 45 to 63 and, if he had done this, the safe would have opened. Thus, according to O’Connor, even though Bill does not know that the combination is 37-45-63, he is able to render it false that the safe is not opened because he is able, in the luck dependent sense, to *ensure* that the safe is opened since he is able to *ensure* that he turns the dial from 37 to 45 to 63. Steve, by contrast, is not able to render it false that Betty does not win the lottery because there is no sequence of bodily movements available to Steve that would ensure that Betty wins: even if he opens the store, she might not pay the fee and, even if she does pay the fee, her number might not be picked. The same holds for Betty: she lacks the luck dependent ability, in O’Connor’s sense, to render it false that she does not win since there is no sequence of bodily movements available to her that would ensure that she wins. If Betty undertakes all of the available bodily movements that put her in a position to win, if she moves her body in the way

³² O’Connor (1993), page 211. O’Connor’s notion is very similar to Ginet’s notion of a broad ability described above.

required to pay the fee, her number may not be picked. Thus, she isn't able to win. Thus, if O'Connor is right, we can conclude that $N(W, t_0)$ and the counterexample is blocked.

I reject O'Connor's reply. Recall, first, that for an agent to be strongly able to do something that would render a proposition P false it is only required that all of the nearest worlds in which the agent exercises one of her abilities, an event e (i) occurs as a causal consequence of her action and (ii) renders P false. Contra O'Connor, it is not required that the agent be able to "directly bring about" event e . In other words, it is not required that there be a sequence of bodily movements available to the agent such that had the agent moved his body in that way, e would have occurred. I claim that if Steve is able to keep the store open he is strongly able to render it false that Betty does not win the lottery. This is so as long as we think in addition that, had he exercised his ability to keep the store open, it might have been a causal consequence of this action that Betty wins the lottery. For if it might have been a consequence of this action that Betty wins, then, by MIGHT HAVE DONE IT, Steve is able to cause to Betty to win the lottery. Applying this to the original BETTY case, Betty is strongly able to render it false that she does not win the lottery so long as we think that, had she exercised her ability to pay the fee, it might have been a causal consequence of this action that she wins the lottery. For if her winning might have been a casual consequence of her paying the fee, then, by MIGHT HAVE DONE IT, she is able to win. Thus, we should not conclude that Betty is not able to win the lottery from the mere fact that she does not possess a way of moving her body that would ensure that she wins.

Even though O'Connor's reply to the BETTY counterexample fails, there is reason to agree with his conclusion. That is, there is reason to think that Betty is unable, at t_0 , to render it false that she does not win the lottery. MIGHT HAVE DONE IT entails that Betty is strongly able to render it false that she does not win the lottery *so long as we think* that, had she exercised her

ability to pay the fee, it might have been a causal consequence of this action that she wins the lottery. However, as I explain next, I believe that this ‘might’-counterfactual is false.

In BETTY, since we know that Betty’s number was not picked and that her number’s being picked is causally isolated from her behavior it is natural to hold that, no matter what Betty does, her number will not be picked and so that, no matter what Betty does, she will not win. That is, the outcome of the lottery would have been exactly the same no matter what Betty had done since the outcome of the lottery is causally isolated from Betty’s behavior. And if it is true that the outcome of the lottery will be exactly the same whether or not Betty pays the fee, then she is not able to win the lottery. Though Betty is able to pay the fee, it is not the case that had she paid the fee, she might have won. Thus, since there is no other act available to Betty such that, had she performed that act, the event of her winning might have occurred, she is strongly unable, at t_0 , to render it false that she does not win the lottery. It follows that $N(W, t_0)$ and so the counterexample is blocked.³³

We might worry that a variation of BETTY poses a problem for the above reply. In the variation, Jim has a choice, at t_1 , concerning whether to let machine A or machine B pick the winning number. Both machines operate indeterministically so that each number in the lottery has an objective chance at winning when the machine is activated. But the random processes in each machine are different. Betty cannot influence the operation of the machines once they are activated, but she would cause Jim to pick machine A by paying the fee. She does not pay the fee and Jim lets machine B pick the winning number. Betty’s number is not picked by machine B.

³³ Many writers on counterfactuals agree with my judgment that Betty would not have won, had she paid the fee. Similar judgments concerning parallel cases are made by Edgington (1995), Kment (2006), and Schaffer (2004).

Again, let V be the proposition that Betty's number is not picked and let W be the proposition that Betty does not win. In the modified case, it is still true that $N(V \rightarrow W, t_0)$. However, unlike the original case, it is not true that $N(W, t_0)$. For Betty is able, at t_0 , to pay the fee and had she paid it, Jim would have let machine A pick the winning number. The result of this alternative random process might have been that Betty wins the lottery. So the MIGHT HAVE DONE IT principle then entails that Betty is able to win the lottery: she is able to pay the fee and had she paid the fee, she might have won. Thus, Betty is able, at t_0 , to do something, namely win the lottery, such that had she done it, the event of her winning would have occurred and rendered W false. If, in addition, $N(V, t_0)$ is true, then we have a counterexample to *BETA*: $N(V, t_0)$ and $N(V \rightarrow W, t_0)$, but $\sim N(W, t_0)$.

However, the modified case does not constitute a counterexample to *BETA* since Betty is able, at t_0 , to render it false her number is not picked ($\sim N(V, t_0)$). The premise $N(V, t_0)$ was true in the original case since, given that the picking of the number is causally isolated from Betty, every act that Betty is able, at t_0 , to perform is such that had she performed that act, her number would not have been picked. In the modified case, however, the picking of the number is no longer causally isolated from Betty. She is able to cause Jim to pick machine A, instead of B, and so she is able to make a difference to which number is picked. Jim's choice of machine A would have been a cause of some number's being picked by machine A. Since Betty's paying the fee would have caused Jim to choose machine A, her payment of the fee would have also been a cause of the number's being picked by machine A. And since her number might have been picked had Jim picked machine A, the MIGHT HAVE DONE IT principle entails that she is able to make it the case that her number is picked: she is able to pay the fee and, had she paid it, she might have made it the case that her number is picked. Thus, since Betty is able to make it

the case that her number is picked, she is strongly able to render V false. Thus, the modified counterexample is blocked.³⁴

Section 5: Quincy

In this section, I argue that David Widerker's (1987) alleged counterexample to *BETA* fails.

Consider QUINCY:

QUINCY: Jim comes to visit Quincy and, at t8, Quincy locks the door, preventing Jim from leaving at t10. Quincy cannot prevent Jim from staying with him at t10 – Jim is able to stay if he chooses to and, if Quincy had not locked the door, he would not have been able to influence Jim's choice to stay in the room or to leave the room. Quincy is able, at t8, to refrain from locking the door.

Here is an argument for the invalidity of *BETA* on the basis of QUINCY. Let P be the true proposition that Jim does not leave by t10. Let Q be the true proposition that Quincy locks the door at t8. Quincy is weakly and strongly unable, at t8, to do something that would render P false since, for anything Quincy is able to do at t8, one of the nearest worlds in which he does it is a world in which Jim does not leave by t10. This is so because no matter what Quincy does, Jim might stay in the room. $P \rightarrow Q$ is true since P and Q are both true. But Quincy is weakly and strongly unable, at t8, to render $P \rightarrow Q$ false since, for anything Quincy is able to do at t8 consistent with the falsity of Q, one of the nearest worlds in which he does it is a world in which Jim leaves by t10 and so is a world in which $P \rightarrow Q$ is true. To put it another way, if Quincy refrains from locking the door at t8, which he must do if $P \rightarrow Q$ is to be false, Jim might leave the room by t10. And Jim's leaving the room by t10 ensures that P is false and so that $P \rightarrow Q$ is true. But Quincy is weakly and strongly able at t8 to do something that would render Q false since he is able at t8 to refrain from locking the door, an event that would explain the falsity of

³⁴ In developing this reply to BETTY and the modified version, I benefited greatly from Kment (2008).

Q. Thus, we have a putative counterexample to *BETA* since $N(P, t_8)$ and $N(P \rightarrow Q, t_8)$, but $\sim N(Q, t_8)$.

QUINCY fails to constitute counterexample to *BETA* for the same reason as BETTY. It is stipulated that Quincy is not able to causally influence Jim's decision to stay in the room or to leave the room. That is, Jim's decision is causally isolated from Quincy's behavior. Thus, had Quincy refrained from locking the door, Jim would not have left the room. Thus, $N(P \rightarrow Q, t_8)$ is false when N is interpreted along the lines of *STRONG*. All of the nearby worlds in which Quincy does not lock the door at t_8 are worlds in which Jim stays in the room at t_{10} . At all of these worlds, $P \rightarrow Q$ is false and the event of Quincy's refraining from locking the door explains the falsity of $P \rightarrow Q$. Thus, Quincy is able to do something, namely refrain from locking the door, such that, had he done it, the refraining event would have (i) occurred, (ii) been Quincy's Xing, and (iii) rendered it false that (Quincy locked the door at t_8 or Jim leaves at t_{10}).³⁵

Section 6: Lewisian Compatibilism

In sections 3 – 5, I considered arguments for the invalidity of *BETA*, arguments that do not presuppose compatibilism. Each of these arguments against *BETA* is consistent with the truth of incompatibilism, as the arguments were stipulated to involve indeterministic situations and were silent about the conditions required for free will. Thus, the arguments, if successful, would demonstrate the invalidity of *B* on the incompatibilist's own terms. I have argued that these arguments fail and so, as far as I can see, there are no arguments against the validity of *B* that do not presuppose compatibilism. What about arguments that do presuppose compatibilism?

³⁵ I noted in footnote 26 that an account the N -operator in terms of the 'might'-counterfactual can handle the COIN case. The same is true for Widerker's case. Consider one of the nearby worlds w in which Quincy does not lock the door at t_8 and in which Jim stays in the room at t_{10} . At w , $P \rightarrow Q$ is false and the event of Quincy's refraining from locking the door explains the falsity of $P \rightarrow Q$. Thus, Quincy is able to do something, namely refrain from locking the door, such that, had he done it, the refraining event might have (i) occurred, (ii) been Quincy's Xing, and (iii) rendered it false that (Quincy locked the door at t_8 or Jim leaves at t_{10}).

6.1 The Standard Lewisian Reply

David Lewis offers the most influential defense of compatibilism. In Lewis (1981), he is responding to a different version of the Consequence Argument than the one under discussion here. The argument Lewis is responding to is a *reductio ad absurdum* argument against compatibilism. I will first explain Lewis' reply to this different argument and then explain how his reasoning can be extended to RCA.

Consider the case ARM:

ARM: Determinism is true and, at t_1 , Jones is deliberating about whether to vote for a proposal by raising his arm at t_2 , but he chooses to refrain from raising it (and so he refrains). The case is non-pathological: Jones' arms are not tied down, he is not paralyzed, he has not been hypnotized to refrain from raising his arm, and so on.

Let H be a proposition describing the entire intrinsic state of the world at some time in the distant past. Let L be the law-stating proposition. Let A be the true proposition that Jones did not raise his arm at t_2 . Suppose, for *reductio*, that Jones is able, at t_1 , to raise his arm at t_2 in ARM. The *reductio* proceeds as follows:

- (1) Jones is able, at t_1 , to raise his arm at t_2 in ARM.
- (2) As a matter of metaphysical necessity, A follows from the conjunction of H and L.
- (3) Had Jones raised his arm at t_2 , A would have been false because Jones raised his arm.
- (4) Jones is able, at t_1 , to do something that would render A false.
- (5) If Jones is able, at t_1 , to do something that would render A false, Jones is able, at t_1 , to render the conjunction of H and L false.
- (6) Thus, Jones is able, at t_1 , to do something that would render the conjunction of H and L false.
- (7) But Jones is not able, at t_1 , to do something that would render H false.
- (8) Thus, Jones is able, at t_1 , to do something that would render L false.

(9) Thus, it is not the case that Jones is able, at t_1 , to raise his arm in ARM.

Premise (2) follows from the definition of determinism. Premise (3) is undeniable, given the sense of ‘because’ introduced in chapter 1. Premise (4) follows from (1) and (3), given the account of the *able to render false* idiom in terms of both WEAK and STRONG. Premise (5) follows from premise (2) and the following general principle:

GENERAL: For all agents S and propositions P and Q, if S is able to render P false and P follows from Q, as a matter of metaphysical necessity, then S is able to render Q false.

Premise (6) follows from (4) and (5). Premise (7) is the Fixity of the Distant Past Premise.

Premise (8) follows from (6) and (7), given the premise that if an agent is able to render (P & Q) false, then, if he is not able to render P false, he is able to render Q false. Since (8) is absurd, we conclude (9) by *reductio*.

The Lewisian reply to the argument begins by disambiguating two senses of the *able to render false* idiom, WEAK and STRONG:

WEAK: An agent has the weak ability, at t, to do something that would render a proposition P false if and only if the agent is able, at t, to do something such that all of the nearest worlds w in which he does it are worlds in which an event e (i) occurs and (ii) renders P false at w.

STRONG: An agent has the strong ability, at t, to do something that would render a proposition P false if and only if the agent is able, at t, to do something X such that all of the nearest worlds w in which he does it are worlds in which an event e (i) occurs, (ii) is the agent’s Xing or an effect of his Xing, and (iii) renders P false at w.

Given WEAK, Lewis accepts (8) and denies that the argument constitutes a *reductio*. Jones is able to do something, namely raise his arm, such that were he to do it, a law-breaking event would have occurred. It would be absurd to suppose that an agent has the ability to perform an action that might be or cause an event that falsifies a law. The compatibilist would be committed to an absurdity if, for example, he was committed to thinking that an agent has the ability to

throw a rock faster than light. But it does not follow from the fact that Jones is weakly able to render L false that Jones is able to perform an action that would itself falsify a law of nature or that would cause an event that falsifies a law of nature. Thus, given WEAK, (8) is not absurd and the *reductio* is blocked.

Given STRONG, Lewis would agree that premise (8) is false, but he would deny GENERAL. Jones is able, at t_1 , to raise his arm and, had he done so, A would have been false because the arm-raising event occurred. Thus, Jones is able, at t_1 , to render A false. But it is not the case that, for anything X that Jones is able to do, had Jones Xed, (H & L) would have been false because of Jones' Xing or because of an effect of his Xing.³⁶ Had Jones raised his arm, for instance, (H & L) would have been false but the conjunction would not have been false *because* the arm-raising event occurred.³⁷

Lewis' reply to the *reductio* argument does not "beg the question" against the incompatibilist. After all, the *reductio* argument functions by drawing an absurd conclusion from the assumption of compatibilism, from the assumption that Jones is able to raise his arm. Given that assumption,

³⁶ It is important to note that my presentation of the Lewisian reply differs from what Lewis actually says in his (1981). This is due to the fact that Lewis is working with a different definition of the *able to render false* idiom than the definition I am working with here. Lewis says that Jones is strongly able to render false the conjunction of H and L. Since Jones is able to raise his arm in ARM, he is able to do something, namely raise his arm, such that necessarily (if he raises it, then $\sim (H \wedge L)$). Given Definition 1 of the able to render false idiom (see chapter 1), the definition Lewis is working with, it follows that Jones is strongly able to render false the conjunction (though he cannot strongly render false either of the conjuncts). So, given Definition 1, Lewis would reject the following implicit premise of the *reductio*: if an agent is able to render (P & Q) false, then, if he is unable to render P false, he is able to render Q false.

³⁷ Benjamin Schnieder (2004) shows that GENERAL is false, given an account of the able to render false idiom like that given by Definition 4. Here is the argument. The false proposition that $2 + 2 = 5$ entails A. Since Jones is able to render A false GENERAL entails that he is able to render it false that $2 + 2 = 5$. But there is nothing that Jones is able to do such that that necessary falsehood is false because he did it. Thus, GENERAL is false. Notice, first, that this case is powerless against the inference rule at play in RCA. A counterexample to BETA requires that we have a true N p and N (if p, then q), but a false Nq, and this requires that p and q are both true. Schnieder's case, though it provides a counterexample to GENERAL, does not satisfy these conditions, i.e. ' $2 + 2 = 5$ ' is false and so 'N ($2 + 2 = 5$)' is false. Second, it seems that we can avoid this problem with the *reductio* argument by modifying GENERAL as follows:

GENERAL*: For all agents S and **true** propositions P and Q, if S is able to render P false and P follows from Q, as a matter of metaphysical necessity, then S is able to render Q false.

Lewis is correct that, given WEAK, premise (8) is not absurd and that, given STRONG, GENERAL is false.

RCA is not a *reductio* argument against compatibilism. But it is clear how we might extend the Lewisian reply to the *reductio* into a reply to RCA. Compatibilists believe that, at t_1 , Jones has the ability to raise his arm at t_2 in ARM. There may be conditions that render an agent unable to do otherwise, but ARM is stipulated to be non-pathological – whatever conditions that the compatibilist takes to deprive an agent of the ability to do otherwise are not present in ARM. Let A be the proposition that Jones does not raise his arm at t_2 . With ‘N’ interpreted along the lines of STRONG, the compatibilist will agree that $N(h(t_0) \& L, t_1)$ and that $N((h(t_0) \& L) \rightarrow A, t_1)$, where $t_0 < t_1$. But the compatibilist denies that $N(A, t_1)$: Jones is able, at t_1 , to raise his arm at t_2 and had he done so, the arm-raising event would have occurred and rendered A false. Thus, the compatibilist holds that perfectly ordinary cases, like ARM, constitute counterexamples to the validity of *BETA*.

The Lewisian reply to RCA, i.e. the argument against the validity of *BETA* on the basis of ARM, presupposes compatibilism. That is, the proposed counterexample to *BETA* presupposes that in ARM, a situation stipulated to be deterministic, Jones is able to do otherwise than he does. As a result of this, the defender of RCA will of course not accept that ARM constitutes a counterexample to *BETA*. The question of whether Jones is able to raise his arm in ARM is precisely the question under dispute and, thus, the incompatibilist is likely to complain that the Lewisian is not entitled to the premise that Jones is able to raise his arm at this stage in the dialectical context.

The Lewisian reply to RCA is certainly dialectically ineffective in a certain way. It has no prayer of convincing anyone not already committed to compatibilism. Nevertheless, the reply

effectively deflects the argument, showing that the compatibilist is not committed to an absurdity.

6.2 Huemer's Principle

Many incompatibilists have not been persuaded by the Lewisian reply to the Consequence Argument. These incompatibilists find it implausible to hold that an agent has the ability to perform a certain action, if the agent's performance of that action depends on the occurrence of an event that is guaranteed not to occur, an event over which the agent also has no control. But this is precisely what Lewisian compatibilism requires. Michael Huemer ((2000) & (2004)) crystallizes this objection to Lewisian compatibilism in a particularly vivid way.

Consider Huemer's Principle:

Huemer's Principle: S is unable, at t, to X if, in order for S to do X, (i) something Y would have to happen that S has no control over at t, and (ii) in fact Y will not (or did not) happen.

Aside from claiming that this principle is intuitively plausible, Huemer holds that the principle offers the most natural explanation of certain cases that everyone, compatibilists and incompatibilists alike, needs to explain. I first explain Huemer's argument for the principle and then I show how the principle can be used to argue, independently of RCA, for incompatibilism. I conclude the section by explaining how the Lewisian compatibilist ought to respond to Huemer's argument.

Consider, first, the case EMERGENCY ROOM (ER):

ER: "Suppose that you are a doctor in the emergency room of a hospital, where a heart attack victim has just been brought in. Suppose you know, as a matter of the laws of biochemistry and the physiology of the human body, that in order for a heart attack patient to be revived, CPR must be administered within three minutes of the time of the heart attack. Suppose you also know that for this patient, four minutes have already elapsed, during which the patient did not receive CPR. You

would obviously be correct to conclude that, at this point, you cannot revive the patient.”³⁸

Everyone agrees that you are unable to save the patient in ER. But *why*, exactly, should we draw this conclusion? Huemer initially claimed that the following explanatory principle is supported by ER:

Past Principle: S is unable, at t, to X if, in order for S to do X something would have to happen before t that in fact does not happen.³⁹

Huemer’s Principle is more general than the Past Principle and it is supported by a variation of ER, ER*:

ER*: “You are a doctor in the emergency room of a hospital, where a heart attack victim has just been brought in. Your hospital has a resuscitation machine which is capable of reviving patients under certain conditions when ordinary (human-applied) resuscitation techniques would not succeed. Suppose you know that, because of the condition of this particular patient and the laws of biochemistry and physics, in order for this patient to be revived, the resuscitation machine would need to be used on him within the next five minutes. Unfortunately, you also know that the machine requires electrical power, that the hospital has suffered a power failure, that the power will not in fact be restored within the next five minutes and you cannot restore the power or do anything that might cause the power to be restored. You conclude that you cannot save the patient.”⁴⁰

In ER, the event that is required for you to save the patient is a past event. But ER* suggests that the *past-ness* of the event is not essential to the case. It is enough if the event is an event that will not occur and that you do not have any control over. A past event, such as the required event in ER, satisfies these conditions: required past events that have not occurred are not under your control. But future events might satisfy these conditions as well, such as the event in ER* of the power being restored.

³⁸ Huemer (2000), page 542.

³⁹ This version of the Past Principle is not exactly the same as the principle in Huemer (2000). The principle in Huemer (2000) is not explicit about the references to time. The above version is, however, implicit in Huemer (2000) and he explicitly puts the principle this way in Huemer (2004).

⁴⁰ Huemer (2004), page 415.

If Huemer's Principle is what underlies our acceptance of the conclusion that you cannot save the patient in ER and ER*, then the Lewisian reply to RCA fails. Given that Jones' raising his arm requires the occurrence of a miracle event, a past event over which Jones has no control at t_1 , Jones is unable, at t_1 , to raise his arm. Thus, not only is Jones strongly unable, at t_1 , to render it false that he does not raise his arm, but he is also weakly unable, at t_1 , to render that proposition false.

Before considering how the Lewisian ought to respond to Huemer's argument, I will note an important confusion in Huemer's presentation of the argument. Consider the following passage concerning the original ER case:

"The nearest possible world in which you save the patient is one in which the electrical power somehow (through none of your doing) gets restored within the next five minutes, leading you then to use the resuscitation machine on him. *The existence of this world is not enough to give you the freedom to revive the patient.* Likewise, it is unclear how the existence of the possible world in which a divergence miracle occurs at t_2 , leading to my later doing B at t_4 , *gives me freedom now to do B at t_4* " (415, italics added).

I have encountered this sort of objection to the Lewisian compatibilist repeatedly in conversation. Whether or not the objection is on to something important, it involves an important confusion about Lewis' position. Lewis' defense of compatibilism does not proceed by offering an account or analysis of the ability to do otherwise. In particular, Lewis certainly does not, and the Lewis-style compatibilist need not, hold that the counterfactual *had the miracle event occurred, Jones would have raised his arm* "gives" Jones freedom. Lewis (1981) does not take the truth of certain counterfactuals to properly analyze the ability to act. So, contra Huemer, the Lewisian compatibilist does not hold that the nearby worlds in which Jones raises his arm constitute or explain the fact that Jones is able to raise his arm. The Lewisian reply is *neutral* as to the correct compatibilist analysis of *agent S is able to perform act X*. What generates the

confusion here is that Lewis' argument distinguishes between the weak and strong senses of the *able to render false* idiom and both senses, as we have seen, are stated in terms of counterfactuals. So there is a quasi-technical notion of ability, the ability to do something that would render a proposition false, which is analyzed in terms of counterfactuals. But the ability to render a proposition false is not the ability to perform an action. Lewis himself never suggests that the ability to render a proposition false, in either sense, should be thought of as an analysis of the ability to perform an action. After all, the *able to render false* idiom is defined in terms of the notion of an agent's having the ability perform an action, a notion that Lewis uses, but never attempts to illuminate.

Now it is certainly true that, if we accept Huemer's Principle, the compatibilist cannot respond to RCA by rejecting *BETA* on the basis of an ordinary case like ARM. However, as Huemer notes, once we accept Huemer's Principle, we have a simple argument, SIMPLE, for the skeptical incompatibilist conclusion:

- (1) S is unable, at t_2 , to do X if, in order for an agent S to do something X at t_2 , an event would have to occur before t_2 , an event that (i) has not happened (and so will not happen) and (ii) that S is unable, at t_2 , to bring about. [Huemer's Principle]
- (2) If determinism is true, then in order for S to do X at t_2 , an event would have to occur before t_2 , an event that in fact has not happened and that S is unable, at t_2 , to bring about.
- (3) Thus, if determinism is true, S is not able, at t_2 , to do X.

We might think that this argument does implicitly invoke the Relative Fixity of the Past premise: that S is unable, at t_2 , to bring about an event prior to t_2 follows from the Relative Fixity of the Past. But the argument makes no appeal to the Relative Fixity of the Laws – it requires only that the laws are deterministic. RCA is unnecessary for establishing incompatibilism, once we have Huemer's Principle in hand. Still, it seems that the compatibilist is in trouble.

Huemer's Principle certainly sounds intuitively plausible on its own and it offers a compelling, natural explanation for your inability to save the patient in ER and ER*.

Lewis would have responded to Huemer's argument in the following way.⁴¹ Huemer's Principle, though it is confirmed by the cases ER and ER*, is false. A narrower principle is available, a principle that justifies the conclusion that you are unable to save the patient in ER and ER*:

Narrow Principle: S is unable, at t_2 , to X if, in order for S to do X, (i) something Y would have to happen and (ii), for any action Z that S is able, at t_2 , to perform, Y will not happen, if S exercises his ability to Z.

In ER, the patient would have to have been brought in at some earlier time in order for you to save the patient. But there is no act that you are able to perform such that, if you performed that act, the patient would have been brought in earlier than he was. Thus, by the Narrow Principle, you are unable to save the patient in ER. In ER*, the power would have to have been restored in order for you to save the patient. Moreover, there is no act that you are able to perform such that, if you had performed that act, the power would have been restored. Thus, by the Narrow Principle, you are unable to save the patient in ER*.

The Narrow Principle blocks Huemer's argument for incompatibilism. The following argument is parallel to SIMPLE, but it utilizes the Narrow Principle instead of Huemer's Principle:

(1') S is unable, at t_2 , to X if, in order for S to do X, (i) something Y would have to happen and (ii), for any action Z that S is able, at t_2 , to perform, Y will not happen, if S exercises his ability to Z.

(2') If determinism is true, then in order for S to do X, (i) a miracle event would have to happen and (ii), for any action Z that S is able to perform, the miracle event will not happen, if S exercises his ability to Z.

⁴¹ My reply here on behalf of the Lewisian is largely derived from an unpublished note on Huemer's Principle from Gideon Rosen.

(3') Thus, if determinism is true, S is unable, at t_2 , to X.

The Lewisian will reject (2'). Return to ARM. Jones will not raise his arm unless a miracle event occurs. But, the Lewisian compatibilist holds that since Jones is able to raise his arm it is not the case that the miracle event will not happen, if Jones exercises his ability to do something that he is able to do. Were Jones to exercise his ability to raise his arm, the miracle event would happen. So, while the Narrow Principle enables the Lewisian to justify the conclusion that you are unable to save the patient in ER and ER*, the principle cannot be used to give a simple argument for incompatibilism.

The Lewisian reply effectively deflects Huemer's challenge. It shows that the compatibilist can hold that Jones is able to raise his arm in ARM, without committing herself to the implausible claim that you are able to save the patient in ER and ER*.

6.3 An Argument for *BETA* and a Lewisian Reply

Van Inwagen (1983) attempts to justify *BETA* along the same lines that Huemer attempted to justify Huemer's Principle. That is, Van Inwagen appeals to confirming instances in which the validity of the principle is allegedly disclosed to us, cases in which *BETA* provides the most natural justification for our reasoning. In this section, I explain Van Inwagen's argument and explain how the Lewisian can respond, a response analogous to that given to Huemer's Principle.

Consider the following case SUN:

SUN: The sun is going to explode in 2015. There is, in principle, a device that somebody could build to prevent the explosion, but nobody would be able to develop the technology required for the device until 2050. If the sun explodes, all life on Earth will die out by 2017. It may, in principle, be possible to sustain life on Earth without the presence of the sun, but this would require technology that nobody would be able to develop until 2050.

Let O be the proposition that there will be no life on Earth in 2017. It is obviously correct to conclude, given SUN, that O is true and that there is nothing that anybody is able, now, to do that

might render O false. Van Inwagen's thought is that what licenses this conclusion is something like principle *BETA*. That is, given the set up in SUN, the most natural way to justify the conclusion is by way of the following argument, where 't₁' picks out the present moment:

N (The sun explodes in 2015, t₁).

N (The sun explodes in 2015 → there is no life on Earth in 2017, t₁).

Therefore, N (There is no life on Earth in 2017, t₁).

The compatibilist, of course, believes that this argument relies on an invalid inference rule. But then how can the compatibilist justify the conclusion that there will be no life on Earth in 2017 and that there is nothing that anybody is able, now, to do that might render that proposition false? What the compatibilist needs is a principle that allows us to draw the required conclusion in SUN, but that does not license the incompatibilist conclusion.

Here is an attempt at identifying a narrower compatibilist-friendly principle. Let 'N_C (P, t)' be equivalent to 'P and nothing anybody is able to do, at t, is such that, were they to do it, P might have been false'. Let 'N (P, t)' be defined as before.⁴² Here is one candidate for the compatibilist's inference rule, *BETA_C*:

BETA_C: $\forall t$ From N_C (P, t) and N_C ((P → Q), t) infer N (Q, t).

The compatibilist can then argue as follows:

N_C (The sun explodes in 2015, t₁).

N_C (The sun explodes in 2015 → there is no life on Earth in 2017, t₁).

Therefore, N (There is no life on Earth in 2017, t₁).

⁴² 'N (P, t)' is equivalent to 'P and nobody is able, at t, to do something that would render P false'.

Can we give an argument structurally similar to RCA in terms of $BETA_C$ instead of $BETA$?

Two central premises of this alternative argument will be the Relative Fixity of the Past_C and the Fixity of the Laws_C⁴³:

Relative Fixity of the Past_C: $\forall t \forall t' (t < t' \rightarrow N_C (h (t), t'))$

Fixity of the Laws_C: $\forall t N_C (L, t)$

The compatibilist will reject both premises in the standard Lewisian manner. Consider Jones in ARM. He is able, at t , to raise his arm and, if he were to raise his arm, $h (t-)$ would have been false, where $t-$ is a time at which the miracle event might have occurred, a time prior to t . Thus, it is not the case that $\forall t \forall t' (t < t' \rightarrow N_C (h (t), t'))$. Moreover, had he raised his arm, L would (and so might) have been false. Thus, it is not the case $\forall t N_C (L, t)$.

The dispute here over $BETA$ is similar to the dispute over Huemer's Principle. In both cases, the incompatibilist offers a principle that he takes to be justified on the basis of instances that confirm the principle. But, in both cases, the compatibilist has an alternative principle available that explains the intuitive conclusions we wish to draw. And the alternative principle, in both cases, blocks the incompatibilist conclusion, given the assumption that, in ARM, Jones is able to raise his arm.

⁴³ Here is the complete alternative argument:

- | | |
|---|---|
| (1) $t_0 < t_1$ | [Stipulation] |
| (2) $\Box \forall t ((h (t) \wedge L) \rightarrow G)$ | [Determinism] |
| (3) $\forall t' \forall t N_C (((h (t) \wedge L) \rightarrow G), t')$ | [2, ALPHA _C : From $\Box P$ infer $\forall t N_C (P, t)$.] |
| (4) $\forall t \forall t' (t < t' \rightarrow N_C (h (t), t'))$ | [Relative Fixity of the Past _C] |
| (5) $N_C (h (t_0), t_1)$ | [1, 4] |
| (6) $\forall t N_C (L, t)$ | [Fixity of the Laws _C] |
| (7) $N_C (L, t_1)$ | [1, 6] |
| (8) $N_C (h (t_0) \wedge L, t_1)$ | [5, 7, N _C -Agglomeration: $\forall t$ From $N_C (P, t)$ and $N_C (Q, t)$ infer $N_C ((P \wedge Q), t)$.] |
| (9) $\forall t N_C (((h (t_0) \wedge L) \rightarrow G), t)$ | [1, 3] |
| (10) $N_C (((h (t_0) \wedge L) \rightarrow G), t_1)$ | [1, 9] |
| (11) $N (G, t_1)$ | [8, 10, BETA _C] |

The incompatibilist may hold that since reasoning with *BETA* appears valid in every case that is not directly at issue in the compatibilism debate, we have reason for thinking that the principle is valid in general. The Lewisian rejection of *BETA* may thus seem to the incompatibilist like an ad hoc, last ditch effort to hold on to compatibilism. But while the compatibilist might well be disappointed that he cannot reply to RCA on the incompatibilist's own terms, the incompatibilist complaint in question oversells his case for *BETA*.

First, it is worth noting that there are many philosophical principles that, though *prima facie* compelling and confirmed by a large number of instances, turn out to be incorrect. The Fregean Comprehension Axiom – there is a set *S* the members of which are just those objects that satisfy the predicate *F* – while it is intuitively compelling and confirmed by many instances, famously leads to contradiction.⁴⁴ Simple statements of global consequentialist principles certainly strike many of us, on first reading, as being correct, even tautologous. But these principles are false nevertheless. Sometimes, as the case of Frege's Comprehension Axiom shows, intuitively compelling, widely confirmed principles are subtly incorrect. And sometimes, as the consequentialism case shows, the principles conflict with our considered judgments about particular cases and are reasonably rejected on this basis.

Second, evidence for a principle derived from confirming instances is significantly weakened when competing principles are available that are confirmed by the same instances. And, as I've argued above, the compatibilist can appeal to competing principles to explain the sorts of cases invoked to support both Huemer's Principle and *BETA*.

⁴⁴ Thanks to Thomas Kelly for this example.

Chapter 3: Two Conceptions of the Free Will Problem and Luck Dependent Abilities

1. Introduction

It is tempting to think that the Consequence Argument has bottomed out in a dialectical stalemate⁴⁵, that there are no available arguments that either party to the dispute can appeal to that ought to rationally compel the opposing party to switch sides. This may be so, but it may be nonetheless be possible to make progress. One promising way for the compatibilist to advance the debate would be to supplement the Lewisian reply to RCA by defending a positive theory of ability, a theory that supports compatibilism. I consider a compatibilist theory of ability that has been used in this way in Chapter 4. But we should first take a closer look at what the free problem is really about. That is the main task for this chapter.

In section 2, I consider Kadri Vihvelin's conception of the free will problem. Vihvelin holds that ability sentences – sentences of the form 'S is able to X when C' – are context sensitive, expressing different propositions on different occasions of use. In particular, predicates of the form 'is able to X' express different properties in different contexts. Given this context sensitivity, it is natural to think that there are many free will problems. For suppose that ability sentences attribute to S an ability₁ in one context, an ability₂ in a second context, an ability₃ in a third context, and so on. It then appears that there is the ability₁ compatibilism dispute, the ability₂ compatibilism dispute, the ability₃ compatibilism dispute, and so on. In other words, for each available sense of 'able', there is the question of whether or not an agent can possess the ability to do otherwise at a deterministic world. But Vihvelin believes that, while ability sentences are context sensitive, we can be sure that ability sentences have the same content in the mouths of compatibilists and incompatibilists in the context of the compatibilism dispute. There are not many free will problems, one problem corresponding to each available sense of 'able'.

⁴⁵ Fischer (1994) describes the dispute in this way.

There is just one compatibilism dispute. For as Vihvelin sees it, the dispute begins with a definitive specification of a context, a specification that thereby ensures that both parties to the dispute are using ability sentences in the same sense.

After explaining Vihvelin's conception of the compatibilism dispute, I argue that it does not isolate a unique problem: the context specified by Vihvelin does not force a unique reading of ability sentences. Vihvelin believes that her contextual specification identifies a form of ability that is *not* the luck dependent form of ability. However, in some common types of situations that satisfy Vihvelin's contextual criteria, ability sentences do in fact attribute luck dependent abilities to agents.

My second aim in this chapter is to examine the notion of luck dependent abilities in more detail. After sketching a plausible framework for analyzing this form of ability in section 3, I argue that, on Vihvelin's conception, the compatibilism dispute over luck dependent abilities reduces to a non-substantive, semantic question. In section 4, I present an alternative conception of the free will problem. According to this alternative, the compatibilism dispute is essentially a normative dispute over whether or not agents at deterministic worlds can have the "abilities" required for moral responsibility.

2. Vihvelin's Conception of the Free Will Problem

Vihvelin (2009) characterizes the compatibilism dispute in roughly the following way. We are often in situations in which we are deliberating, attempting to decide between two actions, A and B. We form a judgment about what to do and then decide to A or decide to B. We then perform the action in question on the basis of this choice. While deliberating, we believe that both actions A and B are open to us in the sense that we are both able to A and able to B. The compatibilism dispute, Vihvelin maintains, is a dispute over whether or not *that* belief, a belief

that we have in the deliberative context to the effect that we are able to A and able to B, can be true at a deterministic world.

Vihvelin holds that this characterization of the dispute rules out certain interpretations of ability sentences. Suppose that sometimes I take the subway to work and that sometimes I drive my car to work. Given that I have the right constellation of intrinsic properties, we can truly say that I am able to drive to work. Since I know how to drive and since I am not currently paralyzed or otherwise physically incapacitated, there is a clear sense in which I currently have what it takes to drive to work. Following Vihvelin, call this the *narrow* ability to drive my car to work, an ability that I possess in virtue of my intrinsic properties⁴⁶. When deliberating about how to get to work, the narrow ability to drive my car to work is *not* the one that I consider. What I am concerned with, from the deliberative standpoint, is getting to work and this requires, not just that my body is functioning normally and that I know how to drive, but also that my environment cooperates with me in certain ways. If the roads are closed, or if my car has been stolen, then there is a sense in which I am not able to drive my car to work, even though I am narrowly able to do so. The kind of ability, then, that we are concerned with in the deliberative context is one that an agent may lack in virtue of certain facts *extrinsic* to the agent, in virtue of certain unfavorable features of his environment. Such unfavorable features of the environment deprive an agent of the *opportunity* to perform a given action that he may be narrowly able to perform. When an agent has the narrow ability to perform an action and the opportunity to perform it, the agent has what Vihvelin calls the *wide* ability to perform that action. Thus, if my car has been stolen or the roads are closed, then I lack the wide ability to drive my car to work in

⁴⁶ I am following Vihvelin when I say that an agent has the narrow ability to do something “by having the right constellation of intrinsic properties”. However, it is worth noting that Vihvelin fixes on the fact that an agent knows how to X to support the claim that the agent is narrowly able to X. However, if knowing how to X is a factive state, e.g. if it essentially involves a relation to a fact or to a way of doing something, then it is not intrinsic.

virtue of the fact that I lack the opportunity to drive. With this terminology in place, we can state Vihvelin's view more precisely as follows: the compatibilism dispute is about whether agents have the wide ability to do otherwise at deterministic worlds.

Note that, typically, when I am deliberating about how to get to work I do not explicitly consider the possibility that there may be features of my environment that would frustrate my attempt to drive. Normally, the roads are not closed. And since I park my car in the garage, there is no good reason for thinking that it has been stolen. In ordinary cases, I simply take it for granted that my environment would cooperate with my attempt to drive to work (and also with my attempt to take the bus to work). In deliberation, I consider the fact that driving is more convenient and the fact that the bus is cheaper. But, normally, nothing else explicitly weighs into my decision about how to get to work. In particular, those features of my circumstances that would undermine my wide ability to drive to work are not on my radar at all.

Now suppose that Vihvelin is right that, in deliberation, we believe that we have the wide ability to perform two different actions. It follows that our belief typically concerns an ability that we might lack due to circumstances that we are not explicitly thinking about at the time. If my car is in fact stolen, then, though I do not consider this possibility in my deliberation, my belief that I am able to drive to work is thereby mistaken.

This point is important for the incompatibilist who shares Vihvelin's conception of the dispute. We, of course, do not know what the laws of nature are or what the entire intrinsic state of the world is at any time. So we do not take these features of our circumstances into account in deliberation. But it does not follow that our deliberative beliefs about what we are able to do are

not sensitive to these features of our circumstances.⁴⁷

Now, while we may agree with Vihvelin that characterizing the compatibilism dispute in terms of the deliberative standpoint ensures that we are not talking about narrow abilities⁴⁸, I am going to argue that we have not thereby fixed a definite sense of ability sentences as they are ordinarily used. To see this, first consider the following case:

TIGHTROPE: Tim is standing on the roof of building A. The only way to get to the roof of building B is to walk across a tightrope. Tim has never tried to walk a tightrope and so he does not have any special tightrope walking skills, e.g. he does not have exceptional balance.

There are two versions of TIGHTROPE to consider. In both versions, Tim is in a deliberative context that satisfies Vihvelin's contextual criteria. That is, he deliberates about whether or not to (choose to) walk the tightrope and decides to perform one of the actions on the basis of the reasons that support so acting.

In one version of the case, there is a bomb on the roof of building B that is going to explode very soon. If the bomb explodes, it will destroy the world. Tim knows that, while he will probably fall from the tightrope and so fail to defuse the bomb, he might succeed if he tries. (We can suppose that Tim is a very good physicist and knows that it is genuinely undetermined whether or not he will succeed in walking across the tightrope when he tries. There is an objective chance, however small, that he will succeed when he tries.) Tim reasons as follows:

⁴⁷ As Fischer (2007) puts the point:

“When I deliberate, I don’t check the total statement of the past and the laws of nature. But it certainly does not follow that I don’t implicitly presuppose that my freedom is the power to extend the actual past, whatever that is, holding fixed the natural laws, whatever they are. Our epistemic limitations imply that we do not worry about compatibility with the past and the laws when we plan for the future; but that is perfectly compatible with a background presupposition that whatever we can do must be connected to the past by a line (holding fixed the natural laws)” (74 – 75).

⁴⁸ It is easy to imagine somewhat bizarre contexts in which an agent is concerned with whether or not she has the narrow ability to perform a particular action X. Suppose that Suzy will receive a reward if, in a deliberative situation, she truly forms the belief that she is narrowly able to X. Suzy might then explicitly consider her abilities, in the narrow sense, and so form the belief that she is narrowly able to X in order to receive the reward.

“If I do not walk across the tightrope and defuse the bomb, then disaster will strike. I will probably fail to walk across if I try, but I am able to walk across the tightrope since I might succeed if I try.” Tim decides to walk across the tightrope partially on the basis of his belief that he is able to do so. Luckily, he succeeds and defuses bomb.

The second version of the case is just like the first version, except that, instead of a bomb on the roof of building B, there is a delicious cookie. Tim is very hungry and he loves cookies. He reasons as follows: “I would love to eat the cookie and, if I was able to walk across the tightrope, I would do so. But, unfortunately, I am unable to walk across the tightrope. So I am going to refrain from trying to walk across the tightrope.” Tim does refrain from walking the tightrope on the basis of his belief that he is not able to walk across.

The only difference between the cases concerns what is at stake in getting to the roof of building B. We can suppose that it is equally likely that Tim succeeds when he tries no matter what his purpose is. That is, the likelihood that Tim succeeds in walking the tightrope does not vary with what is at stake. Given this setup, whatever abilities Tim has in the bomb case he must have also have in the cookie case and whatever abilities Tim lacks in the cookie case he must also lack in the bomb case. I claim that Tim’s reasoning is unobjectionable in both versions of the case. If it is, then there must be a kind of ability, the *reliable* ability to walk the tightrope, which Tim lacks in the cookie case and a kind of ability, the luck dependent ability to walk the tightrope, which Tim possesses in the bomb case. (Since Tim’s abilities are the same in both versions of the case he must also lack the reliable ability to walk the tightrope in the bomb case and he must possess the luck dependent ability in the cookie case.) It follows that focusing on Vihvelin’s deliberative contexts does not fix the content of ability sentences: while Tim possesses the same abilities in both contexts, the sentence ‘I am able to walk across the

tightrope’ as used by Tim expresses a truth in the first deliberative context, but a falsehood in the second.⁴⁹

When Tim walks across the tightrope and defuses the bomb, the ability that he exercises is a luck dependent ability. Importantly, it is not only in far-fetched scenarios like the bomb case that, in deliberation, agents are concerned with whether or not they have a certain luck dependent ability. When the stakes are high enough, we often choose to perform actions on the basis of our belief that we might succeed while recognizing that success is unlikely. We buy lottery tickets and bet on the underdog. Firemen make heroic attempts, against all odds, to save people from burning buildings. And when the stakes are sufficiently high, these actions are often even reasonable.

This concludes the argument that Vihvelin’s conception of the compatibilism dispute does not identify a unique problem.⁵⁰ There is an ability that Tim lacks in both versions of TIGHTROPE, the reliable ability to walk the tightrope, an ability that an expert tightrope walker possesses. There is also an ability that Tim possesses in both versions of TIGHTROPE, the luck dependent ability to walk the tightrope. In the bomb version of the case, Tim deliberates about what to do in the light of his beliefs concerning his luck dependent ability to walk the tightrope. In the cookie version, Tim deliberates about what to do in the light of his beliefs concerning the fact

⁴⁹ I am not saying that the stakes in a context C determine what ‘able’ means in C. I believe that Tim can speak truly by uttering the following speech in the bomb case: “Expert tightrope walkers are able to walk the tightrope, but not me. I’ve never even tried and I don’t know how to do it. So I am not able to walk the tightrope. I might succeed if I try, but a success would be a mere fluke.” I also believe that Tim can speak truly in the cookie case by uttering the following speech: “Sure, I might succeed if I try and so I am able to walk across. But I don’t have a reliable method for walking across. I am not going to try because it isn’t worth the risk.” So in my view Tim can use ‘able to walk the tightrope’ in both contexts to pick out a certain reliable ability or to pick out a certain luck dependent ability.

⁵⁰ It is worth noting that there are still more common uses of ability sentences in deliberative contexts that seem to attribute different forms of abilities to agents. Phil asks Jeremy to go hiking on Monday morning. Jeremy thinks about it and says: “I am not able to go hiking on Monday because I have to go work.” This kind of reasoning is ubiquitous. In deliberation, we frequently rule out courses of action that we may be able to perform, in Vihvelin’s wide sense or in the luck dependent sense, on the basis of the fact that these courses of action are inconsistent with our settled intentions.

that he lacks the reliable ability to walk the tightrope. Thus, focusing on Vihvelin's deliberative contexts does not force a unique interpretation of ability sentences.

Section 3: Luck Dependent Abilities

In this section, I consider luck dependent abilities in more detail. First, I suggest a plausible framework for analyzing the notion of a luck dependent ability. Second, I argue for the surprising conclusion that, even if incompatibilism is true about luck dependent abilities, compatibilism may be true about forms of ability that preclude luck. Third, I argue for the following deflationary conclusion: if we understand the compatibilism dispute over luck dependent abilities in Vihvelin's way, the disagreement between compatibilists and incompatibilists is not substantive, but merely semantic.

3.1 What are Luck Dependent Abilities?

How should we understand ability sentences in the luck dependent sense? A natural account begins with the thought that ability is a kind of possibility.⁵¹ What kind of possibility? If S is able to X, then, plausibly, it must be metaphysically possible that S Xs.⁵² However, metaphysical possibility is not enough: in most contexts at least, the mere metaphysical possibility that an agent Xs does not suffice for the agent's being able to X. While it is metaphysically possible that I walk on water, for example, I am not able to walk on water. If we are inclined to view ability as a kind of possibility, it appears that we need to restrict ourselves to a privileged class of metaphysically possible worlds. Ability, then, is relative possibility, possibility relative to a contextually salient set of facts. Holding fixed certain facts about my body, water, and the laws of nature, it is not metaphysically possible that I walk on water. On

⁵¹ Thanks to Ryan Robinson for helpful discussion. I benefited here from reading an unpublished paper by Robinson on the semantics of ability.

⁵² However, see Jack Spencer (unpublished). Spencer provides an interesting argument for the conclusion that an agent has the ability to X, though his Xing is metaphysically impossible.

the relative possibility account, that is one thing that I might mean by uttering ‘I am not able to walk on water’.

The relative possibility account of ability sentences can be stated as follows:

REL: ‘S is able to X’ expresses, at a context C, the proposition that S’s Xing is compossible with the C-facts obtaining.⁵³

REL is not a plausible account of some common uses of ability sentences. For example, there is the sense in which an expert tightrope walker is, while Tim is not, able to walk across the tightrope. If we are to accommodate this fact on the basis of REL, it must be that, for some contextually salient set of facts C, it is impossible with C that Tim walks the tightrope, but compossible with C that the expert walks it. But it seems that there is no way to specify the relevant set of C-facts in a way that has this consequence. After all, it is compossible with *all* of the facts about Tim’s circumstances that he walks the tightrope, given the assumption that it is compossible with all of the facts that he try to do so. The relative possibility account does, however, provide us with a plausible framework for capturing the luck dependent sense of ability sentences. But it is unclear exactly how it should be implemented.

The *standard incompatibilist treatment* of ability sentences in the luck dependent sense takes the past and the laws to constitute the set of contextually salient facts. This treatment provides an account of ability sentences that captures a particular notion of free will suggested by many writers sympathetic to incompatibilism. Consider, for example, the following passage from Ted Warfield (2003)⁵⁴:

⁵³ Accounts along the lines of REL have been suggested as accounts of ‘can’ by Lewis (1976) and Kratzer (1977).

⁵⁴ The following passages from Fischer and Mele suggest that the incompatibilist endorses the left-to-right direction of the standard incompatibilist treatment:

“Put metaphorically, if an agent can do X, then one could in principle “trace a path” (along which the actual natural laws obtain) between the actual past and his doing X. Imagine various pathways into the future. The point is that the only such pathways that indicate what an agent has it in his power to do are the ones which are extensions of the actual past and along which the natural laws

“Consider a large book containing a list of all the true statements about my world up to the present moment. Imagine further that the book is continuously being updated. As I go about my daily business, what I do is add to the list of truths about my world. My actions add to the actual past in ways consistent with the laws of nature. My actions don’t erase earlier truths. And none of the additions my actions make are additions that involve my doing things in violation of natural laws. This picture strongly suggests that my being free to do something is that my doing so be consistent with the actual past and the laws.”

If S is able to X (if and) only if S’s Xing is compossible with the past and the laws, then, at any deterministic world, S is only able to do those things that he does do. For any action that S does not perform, his performing that action is impossible with the past and the laws. Thus, given the standard incompatibilist treatment, establishing the incompatibilist conclusion does not require a complicated argument like RCA: incompatibilism follows trivially.

It is worth noting that, while the standard incompatibilist treatment is commonly suggested as a plausible incompatibilist account of ability, it renders RCA unsound where we understand the N-operator along the lines of STRONG. Recall that an agent has the strong ability, at t, to do something that would render a proposition P false if and only if the agent is able, at t, to do something X such that all of the nearest worlds w in which he does it are worlds in which an event e (i) occurs, (ii) is the agent’s Xing or an effect of his Xing, and (iii) renders P false at w. Given the standard incompatibilist treatment, the fixity premises are trivially true, but *BETA* is invalid. To see this, return to the BETTY case from chapter 2.⁵⁵ Betty is unable to perform an action that might cause the picking of her number. Thus, she is not able to render it false that her

obtain: these are the paths whose initial segments correspond to the actual world up to the relevant time of the action” (Fischer (1994)).

“Philosophers happy to talk in terms of possible worlds will say that an agent at a world W is [able] to A at t if and only if she...A-s at t in some relevant possible world. One way to see the disagreement between incompatibilists and compatibilists about determinism and being able to do otherwise is as a disagreement about what worlds are relevant. According to incompatibilists, all and only worlds with the same past and natural laws as W are relevant; they hold the past and the laws fixed” (Mele (2003)).

⁵⁵ See Section 4 of Chapter 2.

number is not picked ($N(P)$). And she is not able to perform an action that might cause an event that falsifies the conditional *if her number is not picked, then she does not win* ($N(P \rightarrow Q)$).

However, given the standard incompatibilist treatment, Betty is able to render it false that she does not win. It is stipulated that Betty is able to pay the fee required to win the lottery and, thus, it is compossible with the past and the laws that Betty pays the fee. Moreover, the case involves the stipulation that the outcome of the lottery is genuinely undetermined and causally isolated from any event that Betty is able to cause. Thus, it is compossible with the past and the laws both that Betty pays the fee and that her number is picked. By the lights of the standard incompatibilist treatment, then, Betty is able to win the lottery. So she is able to do something that might render it false that she does not win ($\sim N(Q)$). Thus, given the standard incompatibilist treatment, *BETA* is invalid since $N(P)$ and $N(P \rightarrow Q)$, but $\sim N(Q)$.

I suggest that the incompatibilist deal with this issue by modifying the standard incompatibilist treatment as follows. I will first introduce some terminology. A *basic action* is an action that one performs but not by performing some other action. Running a marathon is not a basic action because it is an action that one performs by putting one foot in front of the other. Breaking a window is something that one does by, say, throwing a rock (or by doing something of the sort) and so it is not a basic action. Basic actions may comprise only certain mental actions, like choosing, or they may also include simple bodily movements. Where X is a basic action, the incompatibilist can endorse the standard incompatibilist treatment.

I next introduce the notion of a *causally isolated proposition P for S* . A proposition P is causally isolated for S just in case (i) P is true and (ii) for any basic action B that S is able to perform, in the sense of the standard incompatibilist treatment, were S to perform B , P would have been true but not because of any action performed by S . The proposition that Betty does

not pay the fee is true, but it is not causally isolated for Betty. For there is a basic action available to Betty, namely choosing to pay the fee, such that it is not the case that had she so chosen, P would have been true. Had she chosen to pay the fee, she might have succeeded. Thus, had she chosen to pay the fee, it might have been false that she did not pay the fee. The proposition that Betty does not win the lottery, however, is causally isolated for Betty. For any basic action B that she is able to perform, had she performed B, it would still have been the case that her number was not picked (but not because she performed B).

With this terminology in place, we provide the *modified incompatibilist treatment* as follows: where X is any non-basic action, S is able to X if and only if S's Xing is compossible with the past, the laws, and the *causally isolated propositions for S*. As we have seen, the true proposition that her number is not picked is causally isolated from Betty. Thus, Betty's winning is impossible with the past, the laws, and the causally isolated propositions for Betty. Given the modified incompatibilist treatment, Betty is unable to win the lottery and the counterexample to *BETA* is blocked.

Returning to the main thread, note that both the standard and modified incompatibilist treatments deliver readings of ability sentences according to which incompatibilism is true. But there are plausible compatibilist treatments of ability sentences available. Jones' raising his arm may be compossible with many sets of contextually salient C-facts, even if it is not compossible with the past and the laws. Where the contextually salient facts are specified by means of any relatively coarse-grained description of Jones' situation – holding it fixed, for example, that Jones is strong enough to raise his arm, that he knows how to raise it, that he isn't paralyzed or compelled not to raise his arm, and so on – it is compossible with those facts that Jones raises his arm.

3.2 Interlude: Luck Dependent Abilities and Reliable Abilities

It is natural to think that, if we utilize the luck dependent sense of ‘able’, RCA would show that no agent has the ability to do otherwise in *any* sense of ‘able’. For RCA would show that, if determinism is true, then no agent is able, no matter how lucky he is, to do otherwise. And if, no matter how lucky an agent is, he is unable to do otherwise, then presumably the agent does not have the ability to do otherwise in *any* sense. The idea is that the notion of a luck dependent ability is the weakest form of ability and that any agent with a stronger form of ability, like a *reliable* ability, must thereby possess the luck dependent form. If RCA succeeds for luck dependent abilities, then it would establish that no agent has free will in any sense.⁵⁶

It is not obvious, however, that if an agent lacks the luck dependent ability to X, then he thereby lacks the ability to X in the sense of ‘able’ that precludes luck, the sense requiring that the agent possess something like a reliable method for Xing. It could be that one of the incompatibilist treatments provides us with the correct account of luck dependent abilities. But it is consistent with this that possession of a reliable ability is a matter of having some compatibilist-friendly form of ability. Possession of a reliable ability may not be a matter of possessing a luck dependent ability *plus* satisfying certain further conditions.

To see how this could be so, consider Saul, the magician described in section 2 of chapter 2. Since Saul is a magician he has the reliable ability to flip heads, walk through walls, and pick the ace of spades from the deck. (He differs from Raul in these respects). A plausible account of

⁵⁶ Vihvelin (1988), commenting on Ginet’s luck dependent notion, appears to be thinking along these lines:

“Ginet’s weak, luck-dependent sense of ‘can’ is...strong enough for the modal argument. For we can’t, no matter how lucky we are, do anything which would or might *casually contribute* to a difference in either the laws or the past. So the premises of the modal argument are all true, given Ginet’s sense of ‘can’ and hence of *N*. If the argument is valid, then no causally determined agent can, even in this weak, luck-dependent sense, cause any proposition to be false. But if all determined agents lack a necessary condition of free will, then they all lack free will” (239).

Saul's possession of these abilities is the following: Saul is disposed to flip heads (or walk through walls or pick the ace of spades) when he tries.⁵⁷ Suppose that this account is correct and make the plausible assumption that having these dispositions is compatible with determinism.⁵⁸ Next suppose that one of the incompatibilist treatments of luck dependent ability sentences is correct. Consider Saul at a deterministic world and suppose that he does not try to flip heads at t. Saul has, at t, the disposition to flip heads (or walk through walls or pick the ace of spades) when he tries: he has the reliable ability, at t, to flip heads. But, by the lights of the incompatibilist treatment, he lacks the luck dependent ability at t to flip heads since the world is deterministic and he does not flip heads. Thus, given these assumptions, it does not follow from the fact that Saul lacks the luck dependent ability to flip heads that he lacks the reliable ability to flip heads.

3.3 A Deflationary Conclusion

In what follows, I will be assuming that REL provides a plausible framework for treating ability sentences in the luck dependent sense, but I will not take a stand on which particular interpretation is correct. In the bomb version of TIGHTROPE, perhaps Tim expresses a proposition along the lines of one of the incompatibilist treatments. Or perhaps he expresses a compatibilist-friendly compossibility claim. Or perhaps the case has not been adequately specified and so it is indeterminate which proposition he expresses. It could be that, for certain ways of filling out the case, Tim expresses a compatibilist-friendly compossibility claim and that, for other ways of filling out the case, Tim expresses an incompatibilist-friendly compossibility claim. I am going to argue that, no matter which interpretation of ability

⁵⁷ See chapter 4 for further discussion of this dispositional account.

⁵⁸ There is no good reason, in general, for supposing that determinism is incompatible with an object's being disposed to do things that it does not actually do. A glass that never breaks at a deterministic world might nevertheless be fragile.

sentences is correct (if any), the relative possibility account threatens to dissolve the compatibilism dispute over luck dependent abilities, given Vihvelin's way of conceptualizing it.

Suppose that we drop the word 'able', in the luck dependent sense, from our vocabulary. Where we would ordinarily say that 'Jones is able to raise his arm', we are now forced to utter sentences of the form 'Jones' raising his arm is compossible with the C-facts'. We can say, for example, that Jones' raising his arm is compossible with the past and the laws. Or we can say, for some compatibilist-friendly specification V of the C-facts, that Jones' raising his arm is compossible with V.

We may be ignorant about which compossibility claim in particular is expressed by 'Tim is able to walk the tightrope' in a given context. But this is purely semantic ignorance. With an account like REL in hand, compatibilists and incompatibilists should agree about the truth-values at a deterministic world of all the plausible candidate propositions expressed by the ability sentence in question, i.e. both parties to the dispute should agree about the truth of each candidate restricted possibility claim. If there is any residual disagreement, it can only be disagreement over the semantic question of whether the sentence in question, as used in a given context, has a meaning that is correctly assigned by one of the incompatibilist treatments or one of the compatibilist treatments. By coming to know that one of our candidate C-facts specifications is *correct* in a given context, whether the specification is incompatibilist-friendly or compatibilist-friendly, it seems that we will not be coming to learn any new non-semantic fact about Tim. We will just be learning that 'Tim is able to walk the tightrope', in that context, expresses a particular proposition, a proposition that we already know to be true or false.⁵⁹

⁵⁹ The method of argument here is similar to Chalmers' (2011) so-called "method of elimination" for resolving verbal disputes. Balaguer (2010) applies a similar method to the phrase 'free will' to argue for a similar conclusion.

If the compatibilism dispute were reduced, in the manner indicated, to semantic questions concerning the meaning of 'able' in ordinary English, it would be somewhat difficult to see why philosophers have been so preoccupied with the free will problem at all. The problem would, of course, be interesting to the semanticist with an interest in the language of ability. But then what explains the relatively unsophisticated semantic theorizing characteristic of the free will literature? If the problem is a semantic one, it would seem that free will philosophers should first become experts in philosophy of language and linguistics. Papers on the compatibilism dispute ought to begin by making explicit the relevant assumptions about the way in which we determine the meaning of the word 'able'. But this is not, at least not typically, what free will philosophers do. Fortunately, however, it is not at all difficult to see why philosophers have been concerned with the free will problem or why free will philosophers are not generally experts in semantic theorizing. For the typical way of conceptualizing the compatibilism dispute in the existing literature is not Vihvelin's way.

4. The Normative View of the Free Will Problem

Rather than deriving from a concern with what 'able' means in English, the free will problem stems from our concern with issues of practical significance to life. To see this, consider first the fact that we take a certain kind of pride in our accomplishments and place great value on our creative expression. The appropriateness of these attitudes, it is natural to think, depends on the fact that we stand in some sort of important action-theoretic relation to our accomplishments and our creative acts. We perceive determinism as a threat in part because it seems that determinism entails that we do not stand in the action-theoretic relation in question and so would show that our ordinary stance towards our accomplishments and creative acts is mistaken. If we have no choice but to succeed, then a certain sort of pride that we find in success can appear unjustified.

If it is never up to us express our creativity in more than one way or not to express it at all, then it can seem unfitting to value such expression in the way that we do.⁶⁰

The most common way of conceptualizing the free will problem draws a connection between our agency and another central issue of practical significance, moral responsibility. As it is often put, the central question is whether or not determinism is compatible with the ability to do otherwise, where ‘able to do otherwise’ is taken to be a quasi-technical term, one expressing a form of “ability” which captures a necessary condition for moral responsibility. On this normative conception, discovering the meaning of ‘able’ in English is potentially orthogonal to our goal of solving the free will problem. It could be, for example, that while ability sentences always express propositions that are captured by the standard incompatibilist treatment, only certain compatibilist “abilities” are required for moral responsibility.⁶¹

In developing the normative view of the free will problem in what follows, I will focus on the case of moral responsibility for bad actions. Being morally responsible for a bad action is a matter of being blameworthy for that action. I will also be supposing that determinism is consistent with the existence of moral facts and, in particular, that agents can perform bad actions at deterministic worlds.

It is important to note that different philosophers operate with different conceptions of blameworthiness. In my view, it is doubtful that any particular account of the nature of blameworthiness is the unique, correct account. All of the proposed accounts in the moral responsibility literature identify phenomena of great importance to human life and it is

⁶⁰ See chapter 6 of Kane (1996) for further discussion.

⁶¹ It is also possible that the following is a constraint on an account of the meaning of ‘able’ in ordinary English: it delivers a sense according to which ‘able’ expresses a form of ability that is necessary for moral responsibility. In that case, the free will philosopher’s term ‘able’ may just be the ordinary English term. But there are other plausible views concerning the semantics of ability language. If, for example, the meaning of ‘able’ is just a function of ordinary usage, then it may well be that the free will philosopher’s term comes apart from the ordinary term in something like the manner indicated.

implausible to think that one particular proposal identifies what moral responsibility *really* is, whereas the others are accounts of mere shadows of the real thing. Even if there is a single, correct account of what moral responsibility is, there is no consensus in the literature about what this account is.

Consider, first, Nomy Arpaly's account of blameworthiness. She says:

“Blame is not primarily something required or prohibited, like punishment, nor even something that can be appropriate or inappropriate, the way that a brave attitude is appropriate for a soldier. It is first and foremost *warranted* or *unwarranted*, the way that my fear of getting a flu shot is warranted only if flu shots are dangerous to me and my admiration for Primo Levi's writing is warranted only if he is in fact a great writer. To hold someone blameworthy is not, in itself, to hold that any course of action is appropriate in regard to him, but rather to hold that a certain attitude toward him is epistemically rational: there was ill will, there was a wrong act, thus blame is warranted. In this way, *on my view, blame is analogous to holding someone to be a bad businessman or a lousy artist*” (Arpaly (2003), emphasis added).

On Arpaly's view, it is sufficient for an agent's being morally responsible that certain negative moral evaluations of the agent are correct in the light of his conduct. We count as blaming an agent when we make judgments that we would express by assertively uttering the sentences ‘Suzy is a jerk’ or ‘Bill is rotten’ or ‘Steve's action was selfish’. When these sorts of negative evaluations are warranted in the light of the agent's conduct the evaluated agent is blameworthy for his action.

It is obvious that these kinds of negative character evaluations are often true. The truth of the proposition that Hitler was evil obviously does not depend on how Hitler became the person that he was. Even if Hitler's rotten character was the result of severe childhood deprivation or the result of aliens tinkering with his brain, it is still true that Hitler was evil. And the truth of the judgment does not depend on whether or not determinism is true. For all we know, determinism is in fact true and the supposition that it is in fact true gives us no reason for thinking that Hitler

was not evil.⁶² Thus, if there is any variety of ability such that being able to do otherwise is required for Arpaly-responsibility, it must be compatible with determinism.

Next consider Scanlon's (2008) account of blameworthiness. He holds that the "conclusion that someone is blameworthy for something he or she has done is a...judgment that the action shows that person to hold attitudes that impair his or her relations with others." Suppose that Mary is babysitting your child and you discover that, unbeknownst to Mary, your child spent two hours in the kitchen playing with knives while you were away. Mary was busy watching television while chatting on the phone with her boyfriend. Mary is clearly Scanlon-blameworthy for neglecting your child. She demonstrates that she has an insufficient degree of concern for your child's welfare and this impairs your relationship with her. It would be appropriate, for example, for you to stop trusting Mary with your child. You would be justified in firing her and telling your friends not to trust her with their children. It is just obvious that blaming Mary in this way can be appropriate if determinism is true. And it is obvious that, even if determinism is true, a third party can correctly judge that it would be appropriate for you to downgrade your relationship with Mary in the manner indicated.⁶³

The two conceptions of blameworthiness just described – Arpaly-blameworthiness and Scanlon-blameworthiness – seem obviously compatible with determinism. In attempting to make sense of the compatibilism dispute on the normative conception, we would do well to isolate forms of moral responsibility that are not obviously compatible with determinism.⁶⁴ If I

⁶² I should note that, as I understand the predicate 'is evil', it does not follow from the fact that someone is evil that they deserve to suffer for their evil acts. 'is evil', as I am using the expression, is roughly equivalent to something like 'is disposed to harm or injure others without good reason'.

⁶³ There might be other ways of downgrading a relationship with somebody that are not obviously compatible with determinism. So there might be certain forms of Scanlon-blame about which compatibilism is not obviously correct. But the Mary case shows that blameworthiness, as Scanlon conceives it, is compatible with determinism, even if every element of Scanlon's package is not so compatible.

⁶⁴ There may be forms of responsibility that are obviously incompatible with determinism. For example, consider the view that S is morally responsible for a bad act just in case S deserves to be subjected to immense and terrible

am right that Scanlon-blameworthiness and Arpaly-blameworthiness are obviously compatible with determinism, then the normative conception of the free will problem should not focus on these on these lightweight notions. If incompatibilism about the ability condition for moral responsibility is to have a chance at being true, we must therefore be operating with a more heavyweight conception of blameworthiness.

One plausible candidate in this connection is the following Strawsonian view.⁶⁵ For an agent to be Strawson-blameworthy for performing a bad action is for it to be appropriate to resent the agent for performing that action, where resentment is understood to constitutively involve the thought that the agent deserves some sort of harm on account of that action.⁶⁶ What is it for resentment to be appropriate? I offer the following necessary condition: resentment is appropriate only if the cognitive components constitutive of it are true. Thus, an agent is Strawson-blameworthy for a bad action only if the agent deserves to be harmed for performing that action. Since Strawson-blameworthiness is not obviously compatible with determinism, understanding the compatibilism dispute in terms of it appears promising.⁶⁷

torture for all of eternity. You might think it is just obvious that nobody can deserve to suffer in that way, regardless of whether or not determinism is true. We should not understand the compatibilism dispute in terms of this hellish notion, if we find it obviously impossible for an agent to deserve such eternal torture.

⁶⁵ See Strawson (1962) and Rosen (unpublished).

⁶⁶ Compare Hume (1739): “When I receive any injury from another, I often feel a violent passion of resentment, which makes me desire his evil and punishment” (2.3.3.10).

⁶⁷ In light of the apparent widespread ambiguity (or indeterminacy) concerning the term ‘moral responsibility’, I find it baffling that the literature on free will is filled with papers in which the authors use the term without ever saying what they mean. Every paper on so-called “Frankfurt-cases”, for example, requires the reader to make a judgment about whether an agent is morally responsible for an action in a particular case. But in nearly every such paper it is unclear what the author has in mind by invoking the notion of moral responsibility. It may very well be that, for any given Frankfurt-case, an agent is morally responsible in some senses, but not in others. I conjecture that this shortcoming in the literature has led many free will philosophers to engage in pointless disagreement: philosopher 1 and philosopher 2 might agree that agents at deterministic worlds are Arpaly-blameworthy, Scanlon-blameworthy, and not Strawson-blameworthy, while having different forms of blameworthiness in mind in the context of the dispute.

4. Conclusion

I have argued that Vihvelin's approach to the compatibilism dispute fails to isolate a unique problem. In different deliberative contexts, we deliberate in the light of beliefs concerning different forms of ability. In many contexts, we are concerned, in particular, with whether or not we have certain luck dependent abilities. I have sketched a plausible framework for treating the notion of luck dependent abilities, the relative possibility account. I argued that, given this account, Vihvelin's conception of the free will problem reduces to a semantic dispute. I then sketched an alternative conception of the free will problem, one that ties the notion of ability to our concern with moral responsibility. I disambiguated three forms of moral responsibility – Arpaly-blameworthiness, Scanlon-blameworthiness, and Strawson-blameworthiness – and suggested that, to give the incompatibilist a chance, we ought to characterize the dispute in terms of Strawson-blameworthiness rather in terms of the alternative, lightweight varieties of responsibility.

Chapter 4: The Dispositional Theory of Ability

I. Introduction

I concluded chapter 3 by sketching the normative conception of the free will problem. I suggested that Strawson-blameworthiness provides us with one conception of moral responsibility in terms of which we can understand the compatibilism dispute without trivializing it. With this normative conception of the dispute in view, my aim in this chapter is to examine a leading compatibilist conception of abilities, the dispositional theory.

Michael Fara, Kadri Vihvelin, and Michael Smith each endorse the view that the abilities relevant to moral responsibility are to be understood as a particular kind of disposition.^{68,69} The dispositional theories of ability, although similar in certain respects, are importantly different. Fara and Vihvelin each hold that having the ability to act otherwise is a matter of having a certain disposition.⁷⁰ Fara holds, in addition, that this ability is a necessary condition for moral responsibility. Vihvelin rejects this latter claim, holding instead, with Smith, that it is the ability to choose on the basis of reasons that is necessary for moral responsibility.⁷¹

I begin by explaining the motivation for the dispositional theory of the ability to act in section 2. I then consider Fara's account and I argue that it is unsatisfactory. The ability to act otherwise, on Fara's understanding, is not a necessary condition for moral responsibility. After

⁶⁸ Fara, Vihvelin, and Smith each say that certain kinds of dispositional abilities are required for moral responsibility. However, none of them say what they mean by 'moral responsibility'. It could be that they each think that Strawson-blameworthiness is impossible. But I will proceed on the assumption that they think agents may be Strawson-blameworthy for their actions and that dispositional abilities are necessary conditions on Strawson-blameworthiness.

⁶⁹ To my knowledge, these are the only three theorists who explicitly endorse dispositional theories of normatively relevant abilities. However, there are other compatibilist theorists that are naturally understood along similar lines. Jay Wallace, for example, holds that to be morally responsible and agent must have "the capacities of reflective self-control". Wallace does not offer a metaphysical account of capacities, but a dispositional theory of Wallace's capacities appears promising.

⁷⁰ Smith does not offer an account of the ability to perform an action.

⁷¹ 'The ability to choose for reasons' is a quasi-technical term and I unpack it in section 3.

rejecting Fara's account, I turn to Smith's. Though Smith's account is more plausible, I argue that it faces serious problems.

2. The Motivation for the Dispositional Theory

Consider the thesis *the ability to act is a disposition* (AA):

AA: For some suitable value of C, S is able to X if and only if S is disposed [to X when C],

where 'X' is to be replaced by verb phrases expressing voluntary actions that an agent does by trying to do them, phrases like 'run around the building' or 'break the window'. AA is therefore silent about the ability to try to X and the ability to choose to X since trying and choosing, while actions, are not actions that an agent does by trying to do them.

There are several considerations in favor of the view that the ability to act is a disposition. Just as an object O can possess the disposition D to X in C when D is not being manifested, and even if O *never* manifests D, an agent S can have the ability A to X in C when S is not Xing, and even if S never exercises A.⁷² When an object has the disposition or ability to X in C it follows that the object *can* X in C in virtue of the fact that it possesses the relevant disposition or ability. Both dispositions and abilities are naturally thought of as causal powers of objects to produce certain effects. If we understand abilities in terms of dispositions, we have an explanation for these similarities.⁷³

⁷² Expressions like 'in C' and 'when C' can function in two different ways in sentences of the form 'S is able to X in C' or 'S is able to X when C'. On one reading, the sentences tell us when S possesses a particular ability, the ability to X. We might force this reading by using a sentence of the form 'In C, S is able to X' instead. On the second reading, 'in C' and 'when C' do not function by telling us when S possess the ability, but instead modify the thing S is said to be able to do. On the second reading, Xing is not the action that S is said to be able to perform. Rather, it is *Xing when C* that S is able to perform. I adopt the underlining convention in the text to eliminate the ambiguity between the two readings whenever there is a risk of misinterpretation. The same general point applies to disposition ascriptions and I adopt the underlining convention there as well.

⁷³ These points are all made in Vihvelin (1994).

One historically influential implementation of the idea that abilities should be accounted for in terms of dispositions took for granted a particular analysis of dispositions, the simple conditional analysis (SCD):

SCD: An object O is disposed to X in C if and only if O would X, were O placed in C.

An account of abilities directly modeled on SCD fails, i.e. substituting ‘has the ability’ for ‘is disposed’ generates an incorrect account. For an agent might possess the ability to X in C, even though he does not or would not X when in C in virtue of the fact that were he placed in C, he would not try to exercise the ability. Woods, for example, has the ability to sink the putt when he is standing five feet away from the hole, but he will not sink the putt if he does not attempt it. Examples like this suggest that, if our account of abilities is to be modeled on SCD, a notion like *trying* or *choosing* must be somehow incorporated. The simple conditional analysis of ability (SCA), an account directly modeled on SCD does precisely this:

SCA: An agent S has the ability to X in C if and only if, S would X, were S to try to X in C.⁷⁴

Fara’s thesis, FA, incorporates the notion of trying into the relevant disposition ascription, though Fara rejects the simple conditional analysis of dispositions:

FA: S is able to X if and only if S is disposed to X when he tries.

SCA is supported by a wide range of cases. To take an example from G.E. Moore, a certain cat will climb the tree when she tries, whereas a dog will fail to climb the tree no matter how hard he tries.⁷⁵ Or compare a man sitting in his room on his own accord with another man tied to

⁷⁴ I am probably (following Vihvelin (2004)) stretching quite a bit as far as historical interpretation goes. Did Hobbes, who held something like (SCA), first think that abilities are dispositions and then truck in something like (SCD) in presenting his account? Regardless of whether the reasoning presented correctly describes the origins of (SCA) and similar accounts, it is *useful* to think of (SCA) as involving two components: (i) the claim that abilities are dispositions and (ii) a certain implementation of that idea. Thinking of (SCA) this way will prevent the mistaken thought that the success of a dispositional account is essentially tied to the success of (SCA).

⁷⁵ Moore (1912/2005).

a chair – the former man would go for a walk if he tried, while the latter man’s attempts to walk would be frustrated by the chains. SCA gets the intuitively correct result in these cases – the cat is able to climb the tree, while the dog cannot, and the unchained man is able to go for a walk, while the man chained to his chair cannot.

Despite the success of SCA in these sorts of cases, there are good reasons for thinking that SCA fails to provide an adequate account of ability. First, cases of *masked abilities* and cases of *finkish abilities* show that the truth of the counterfactual on the right-hand side of SCA is not necessary for the truth of the corresponding ability ascription on the left-hand side. An agent’s ability to X in C is *masked* just in case the agent tries to exercise the ability, i.e. the agent tries to X, (ii) C obtains, (iii) the agent fails to X, and (iv) the agent does not lose her ability to X. Suppose that Woods tries, but fails, to sink the putt from five feet away. There may have been a sudden gust of wind, or Woods might have momentarily lost concentration, or a sorcerer may have caused the ball to veer just to the right of the hole. In these cases, it is natural to think that Woods still had the ability to sink the putt when standing five feet away from the hole, even though he missed it on this occasion. If so, then while he retained his ability even when he missed the putt, the ability was masked by the gust of wind (or his loss of concentration or the actions of the sorcerer). But SCA has the consequence that Woods lacked the ability to sink the putt when standing five feet away from the hole: it is not the case that had he tried to sink it, he would have since Woods tried to sink it and missed.

An agent’s ability to X in C is *finkish* just in case (i) the agent possesses the ability but (ii) were she to try to exercise the ability, she would lose it, and so thereby fail to X. Woods is measuring his shot on the green. He is able to sink the putt, but a sorcerer, who wants Woods to miss the putt, is waiting in the wings. If and when Woods makes an attempt at sinking the putt,

the sorcerer will change Woods' physical constitution in such a way that he no longer knows how to swing the club with appropriate force or so that Woods is no longer strong enough to even hold the club. But Woods, as he is measuring the shot, knows how to swing the club and is in possession of all the physical capacities required for sinking the putt. So, intuitively, Woods has the ability to sink the putt, even though, due to the sorcerer's presence, Woods would lose the ability to sink the putt if he tried and so he would not sink it if he tried.

The second set of problem cases for SCA, cases of *fluky success* and cases of *finkish lack of ability*, show that the truth of the counterfactual on the left-hand side of SCA is not sufficient for the truth of the ability ascription on the right-hand side. An agent's successful Xing is *fluky* just in case, though she Xs, her Xing was just a matter of luck and so intuitively not an exercise of ability. I am not a magician. I might pick your card out of the deck if I try, but I lack a reliable method for doing so. Suppose that I do pick your card. Since it is true that *had I tried, I would have succeeded* – I did try and did succeed – SCA entails that I have the ability to pick your card. But my success was lucky and so not an exercise of the ability to pick your card.⁷⁶

An agent *finkishly lacks* the ability to X in C just in case (i) she lacks the ability but (ii) she would acquire the ability were she to try to X in C. Harry is a bungling hit man and he is supposed to shoot Smith from 600 yards with a sniper rifle. But he simply cannot keep his hand steady enough to take an accurate shot. A sorcerer, who wants Harry to succeed in shooting Smith, is prepared to turn Harry into an expert marksman, if and when he tries to shoot Smith. Of course, Harry is not currently an expert marksman. He does not have the ability to shoot Smith from 600 yards, but he would succeed if he tried. He would succeed because, due to the

⁷⁶ The dispositional theory is partially motivated by the thought that cases of fluky success do not amount to the exercise of ability. (Consider Smith's (2003) case of Fluky John). So the dispositional theory is not attempting to give an account of luck dependent abilities of the sort discussed in chapters 2 and 3.

presence of the sorcerer, Harry would acquire the ability to shoot Smith from 600 yards if he tried to shoot Smith.

The cases just described – cases of fluky success, and cases of masked, finkish, and finkish lack of ability – are not problems for the thesis that abilities are to be understood in terms of dispositions, but merely problems for the implementation of that thesis provided by SCA. In fact, the existence of the problem cases is actually strong evidence in favor of FA, Fara's central dispositionalist thesis concerning the ability to act. After all, the problem cases appear perfectly analogous to cases that undermine SCD, the simple conditional analysis of dispositions. A glass that is wrapped in protective packaging has the disposition to break when dropped, even though it would not break when dropped due to the packaging. The packaging masks the disposition.⁷⁷ A certain live wire is disposed to conduct electricity. But were the wire put in contact with an object, it would become dead and so fail to conduct. The wire's disposition to conduct electricity is finkish. A certain dead wire is not disposed to conduct electricity. But were the wire put in contact with an object, it would become live. The wire finkishly lacks the disposition to conduct electricity.⁷⁸ A boulder is not disposed to shatter when dropped, but it has a weak spot, an *Achilles Heel*. Were the boulder dropped from precisely 10 meters, it would shatter. As it happens, the boulder was dropped from precisely 10 meters and it shattered.⁷⁹ Since dispositions can themselves be masked, finked, and finkishly lacked, and since objects can have Achilles Heel's, the thesis that abilities are to be understood in terms of dispositions actually *predicts* the existence of the problem cases. The counterexamples to SCA thus suggest that the principle went wrong, not by holding that abilities are dispositions, but by invoking the simple counterfactual analysis of dispositions.

⁷⁷ Johnston (1992).

⁷⁸ Martin (1994).

⁷⁹ Manley and Wasserman (2008).

3. A Preliminary Argument against Normative Dispositionalism

In examining the dispositional theory of ability, I will be focusing on *deliberative situations*. An agent is in a deliberative situation just in case (i) the agent is deliberating between two (or more) courses of action, (ii) one of the actions he is considering performing is the action he ought to perform and all other actions are actions he ought not to perform, (iii) the agent forms a judgment about which action he ought to perform, (iv) he *wills* to perform one of the actions (in accordance with his judgment or not), and (v) his *willing* causes his performance of the action in the normal, non-deviant sort of way.

I am not presupposing a sophisticated account of deliberation or of the notion that one ought to do something. The notion of *willing* is intended to be neutral as to the state that initiates action at the conclusion of deliberation. Some theorists construe willing in passive terms: it is, say, the state of desiring the relevant action. Other theorists construe willing in active terms: it is, say, the act of choosing the relevant action.

The ability to choose to X on the basis of reasons is a complex ability that we typically suppose an agent possesses in deliberative situations in which he is deliberating about whether to X or not to X. I offer the following (partial) analysis: S has the ability to choose to X on the basis of reasons, in a deliberative situation, just in case (i) X is the action S ought to perform, (ii) S has the ability to recognize the reasons for and against Xing, (iii) S has the ability to form the correct judgment about what he ought to do, and (iv) S has the ability to *will* the action that accords with his judgment about what he ought to do.

The ability to recognize reasons is, in part, an active ability and, in part, a passive one. In deliberation, recognizing the reasons for and against ones contemplated courses of action is often a matter of *doing* certain things, of acting in certain ways – performing calculations, focusing

ones attention, and making inferences. But, once the relevant actions have been performed, believing that a consideration is or is not a reason in favor of an action is a passive matter. Forming judgments about what one ought to do is also a passive matter, in this sense. An agent might, in some special circumstances, perform certain actions with the goal of coming to believe something. But, in the usual case, the formation of such a judgment is something that just happens. This is not to say that the judgment is not caused, but only that the judgment is not an action.

In the rest of this section I will give an argument against *normative FA-dispositionalism*, the conjunction of Fara's thesis, FA, and PAP-Failure:

FA: S is able to X if and only if S is disposed to X when he tries.

PAP-FAILURE: In deliberative situations, the ability to X is a necessary condition on being morally responsible for failing to X.

The following case, FLUKE, provides a counterexample to normative FA-dispositionalism:

FLUKE: Tim is standing on the roof of building A. There is a bomb on the roof of building B and it will destroy the world if it explodes. The only way to get to the roof of building B is to walk across a tightrope. Tim has never tried to walk a tightrope and so he does not have any special tightrope walking skills, e.g. he does not have exceptional balance. He does not have the disposition to walk the tightrope when he chooses (tries). Tim, however, truly and reasonably believes that were he to choose to walk the tightrope, he would succeed. He is able to choose to walk the tightrope on the basis of the reasons that support so acting, i.e. in order to defuse the bomb. But Tim decides not to attempt to walk the tightrope and so he does not try to do so. The bomb explodes and the world is destroyed.

Here are two different ways of filling out the details of FLUKE. In the first version of the case, a sorcerer wants Tim to defuse the bomb and Tim knows that the sorcerer would turn him into an expert tightrope walker once he made the decision (or began to try) to walk across the tightrope. On this version of FLUKE, Tim finkishly lacks the ability to walk the tightrope: he is not able to walk the tightrope, but he would acquire the ability if and when he tried to defuse the

bomb.⁸⁰ In the second version of the case, no sorcerer is present, but Tim has an *Achilles Heel*: given that his external circumstances are *just so*, Tim is disposed to walk the tightrope when he tries and circumstances are *just so*. In fact, given a complete specification of his circumstances, Tim is guaranteed to walk the tightrope successfully when he tries.⁸¹

FLUKE is a possible case on any version of the dispositional theory of the ability to act that allows that an agent might lack the disposition to X when he tries, even though he truly and reasonably believes that he would succeed if he tried. PAP-Failure entails that Tim is not morally responsible for failing to walk the tightrope or for failing to defuse the bomb. For, by the lights of FA, Tim lacks the ability to walk the tightrope and so lacks the ability to prevent the explosion. But, intuitively, Tim may be morally responsible for failing to walk the tightrope and for failing to prevent the explosion. It is stipulated that Tim had the ability to choose to walk the tightrope on the basis of the reasons that support so acting. And he truly and reasonably believed that, if he chose to walk the tightrope, he would succeed. It thus seems clear that, if Tim is blameworthy for failing to choose to walk the tightrope in order to defuse the bomb, he is blameworthy for failing to walk the tightrope and for failing to defuse the bomb. It is irrelevant that, had Tim chosen to walk the tightrope and then defused the bomb, his success would have been a mere fluke (or due to his acquisition of certain abilities). Since Tim may be morally responsible for failing to walk the tightrope and for failing to defuse the bomb, while lacking the

⁸⁰ Suppose that something disastrous will happen if a particular dead wire – a wire that is not disposed to conduct when touched – does not conduct electricity. Eli knows this and he knows that a sorcerer will render the dead wire live if and when it is touched. Eli is able to choose to put the dead wire in contact with an object in order to avoid the disastrous consequence, but he does not do so. The sorcerer-version of FLUKE is perfectly analogous to the case of Eli and the dead wire. Since Eli cannot excuse himself by noting that the wire is dead and so not disposed to conduct electricity when touched, Tim cannot excuse himself by noting that he lacked the disposition to walk the tightrope when he tries.

⁸¹ Suppose that something disastrous will happen if a particular brick does not break. Bill knows this and he knows that the brick has an Achilles heel: it will definitely break if dropped *just so* and Bill is able to ensure that the brick is dropped in just that way. He is able to choose to drop the brick in just the right way so as to avoid the disastrous consequence. But he doesn't choose this. The case of Bill and the brick is similar to the second version of FLUKE. Just as Bill cannot excuse himself by noting that the brick lacked the disposition to break when dropped, Tim cannot excuse himself by noting that he lacked the disposition to walk the tightrope when he tries.

ability to do these things by the lights of FA, we must reject either FA or PAP-Failure. Thus, normative FA-dispositionalism is false.

4. A Reply to the Preliminary Argument

The preliminary argument on the basis of FLUKE appealed to the fact that an agent might possess the ability to choose to X on the basis of reasons, while lacking the ability to X. Now suppose that S is not morally responsible for failing to choose to X on the basis of reasons. Corresponding to the three components offered in my analysis of this complex ability, there are three relevant ways in which this could happen. First, S might have blamelessly failed to recognize the reasons that favor Xing. Second, S might have blamelessly failed to form the correct judgment about what he ought to do. Third, S might have blamelessly failed to will to X. It is plausible to think that, were any of these three possibilities realized, S would not have been morally responsible for failing to X in virtue of the fact that he was not morally responsible for failing to choose to X on the basis of reasons. If so, then S's moral responsibility for failing to X is an instance of derivative moral responsibility; S's responsibility for failing to X is explained by his responsibility for something else.

The dispositional theorist can exploit this fact. She can reject PAP-FAILURE and instead endorse PAP-FAILURE*:

PAP-FAILURE*: In deliberative situations, an agent is non-derivatively morally responsible for failing to X only if the agent is able to X.

The argument against normative FA-dispositionalism on the basis of FLUKE fails to impugn *non-derivative* normative FA-dispositionalism, the conjunction of FA and PAP-FAILURE*.

While Tim may be morally responsible for failing walk the tightrope in FLUKE, his responsibility for failing to X is derivative, explained by his responsibility for failing to choose to walk the tightrope in order to defuse the bomb.

5. Smith's Account of the Ability to Choose on the Basis of Reasons

Note that FA cannot constitute a complete account of the ability to choose to X on the basis of reasons. FA concerns only the ability to perform those actions that one does by trying to do them. Thus, neither the ability to see that a consideration is a reason nor the ability to correctly believe that one ought to perform a certain action are covered by FA since both *seeing that a consideration is a reason* and *believing that one ought to do something* are not actions. Moreover, AA does not cover the ability to will to X. Willing is either an act like choosing, something that one does not do by trying to do it, or it is a state like desiring, which is not an action at all. If the dispositional theorist replies to the preliminary argument in the manner indicated, he owes us a dispositional account of the ability to choose on the basis of reasons.

To my knowledge, Smith's (2003) account is the only explicit dispositional account available of the ability to choose on the basis of reasons.⁸² In presenting and evaluating Smith's account, I will focus on one aspect of the ability to choose to X on the basis of reasons, namely the ability to form correct normative judgments, though similar remarks extend to Smith's account of the ability to desire in accordance with ones' normative judgments.

I will develop Smith's view with reference to the following case⁸³:

RECKLESS: A certain reckless woman ought not to have another drink because, if she does, she will neglect her children. She deliberates about whether to have another drink and forms the mistaken belief that she ought to have another. She then acts in accordance with this belief and ends up neglecting her children.

⁸² Vihvelin (1994) also maintains that the abilities involved in the ability to choose on the basis of reasons are the abilities relevant to moral responsibility. However, Vihvelin's analysis does not cover these abilities. Like Fara, her analysis of abilities has *trying* as the stimulus condition in the relevant dispositions. She does not, like Fara does, explicitly restrict 'X' in the analyzandum to phrases expressing actions that one does by trying to do them. So she could think that her theory is an adequate account of the ability to form correct normative judgments. But it is implausible to think that willing an action is something that an agent does by trying to do it and so Vihvelin's analysis is not a plausible account of the ability to will. Vihvelin (personal communication) recognizes this shortcoming of her (1994) account and she has a more sophisticated account in hand, though the account is yet to be published.

⁸³ Smith uses essentially the same case in developing his position.

It is plausible to think that if the reckless woman is not blameworthy for her belief failure, then she is not blameworthy at all. Smith agrees. On his view, for the reckless woman to be blameworthy for neglecting her children she must be blameworthy for failing to believe that she ought not to have another drink.⁸⁴

Determining whether or not the reckless woman is blameworthy for her belief failure is, according to Smith, a two-stage process. At stage 1 of responsibility assessment, we ask whether she had the ability to form the correct normative belief while deliberating.⁸⁵ If she lacks this ability, then she is not blameworthy. If she has the ability, then we move to the second stage of responsibility assessment. At the second stage, we ask whether it was the reckless woman's failure to exercise her ability to form the correct normative belief that is the "relevant explanation" of her belief failure. If it was, then she is blameworthy for her belief failure. If it is not the relevant explanation of her belief failure, then she is not blameworthy.

It will be helpful to consider two versions of RECKLESS in which the woman passes stage 1 but fails to pass stage 2 of responsibility assessment. In these versions of the case, while the woman has the relevant ability to form the correct belief, her failure to exercise this ability is not, according to Smith, the relevant explanation for her belief failure. In the first case, Black, a nefarious neuroscientist, causes the woman to form the incorrect belief. On Smith's view, though the woman may have the ability to form the correct belief, she is not blameworthy. For the relevant explanation of her belief failure, Smith holds, is not her failure to exercise the ability

⁸⁴ Smith says: "what we blame her for is having the wrong belief about what she should do in the circumstances of action that she faces." He never explicitly says that she may be blameworthy for the *actions* in question. I am supposing that Smith believes, in addition, that the reckless woman may be blameworthy for her actions. He focuses on her blameworthiness for her belief failure because this is the primary locus of responsibility; her responsibility for her actions is grounded in her responsibility for her belief failure.

⁸⁵ Smith actually uses the term 'capacity' where I am using the term 'ability'. I do not think that anything of significance to my presentation of Smith's view hangs on this terminological deviance. Smith also sometimes expresses the relevant points in terms of what an agent *can* or *cannot* do. 'Can', in some contexts, is certainly used to attribute abilities (capacities) to agents, but I have systematically avoided using 'can' and cognate expressions because they are even more systematically ambiguous (or context-sensitive) than 'able' and its cognates.

to believe correctly, but is instead the behavior of Black. There are also conditions internal to the agent that might undermine responsibility in the way that Black does. In the second version of the case, the woman possesses the ability to believe correctly, but she is blamelessly filled with self-hatred that she can neither acknowledge nor eradicate. Her being in that state explains her belief failure. In such a case, Smith holds that the woman is not blameworthy for her belief failure since the relevant explanation of her failure is not that she failed to exercise the ability to believe correctly. Instead, the relevant explanation is that she is filled with self-hatred.

How exactly do we make the stage 1 and stage 2 responsibility assessments? In other words, how do we determine whether an agent has the ability to form the correct belief while deliberating and how do we determine what counts as the relevant explanation for an agent's belief failure? Begin with stage 1. Smith maintains that the reckless woman has the ability to believe correctly just in case a certain "raft of counterfactuals" is true of her. We consider a range of situations in which the woman deliberates about what she ought to do, situations that, while similar to her actual situation, differ from actuality in any of a number of small ways. In some of these situations, for example, the temperature in the room will differ slightly from how it actually is. Or the woman might pose the question *what should I do?* to herself in slightly different ways, with slightly different wording or with a different tone of voice or at a slightly different time. If the woman forms the correct belief in a suitable proportion of these alternative situations, then she is able to form the correct belief. If she fails to form the correct belief in enough of these situations, then she is unable to form the correct belief.

What is the woman like at the outset of these possible deliberative situations? As Smith sees it, the woman might differ from the way that she actually is in certain extrinsic and intrinsic respects. To see why, consider a case presented by Smith, the case of Mark Johnston's shy and

intuitively powerful chameleon.⁸⁶ The chameleon is sitting in the dark and is not being viewed. The chameleon is green. But the chameleon, due to its intuitive power, anticipates when it is about to be viewed and then, due to its shyness, blushes bright red. Smith maintains that the chameleon is disposed to look green when viewed. What must the chameleon be like at the outset of the possible viewing situations that are relevant to the assessment of its disposition to look a certain way when viewed? If we consider viewing situations in which the chameleon has retained *all* of its intrinsic properties, then we will presumably be forced to say that the chameleon is disposed to look red when viewed: he will intuit that the viewing is imminent and will thus blush red before he is viewed. The lesson to be drawn from the case, Smith holds, is that the disposition to look green when viewed is a constituted disposition, a disposition that an object has in virtue of a subset of its intrinsic properties, in particular, in virtue of certain intrinsic properties of its surface. The viewing situations that are relevant to the chameleon's disposition to look a certain way when viewed involve a chameleon that retains these surface properties. But we "abstract away" from any other properties of the chameleon that might have an effect on its color. In particular, we abstract away from its shy and intuitively powerful psychology.⁸⁷

⁸⁶ Johnston (1992).

⁸⁷ Smith's account of dispositions is developed along the lines of Johnston's (1992) account. But whereas Johnston appealed to a single, special counterfactual, Smith suggests that we consider a wide range of counterfactuals. In assessing whether O is disposed to M when C, Johnston's account asks us to consider a counterfactual like this: were O in C, and were there no masking conditions present, O would M. Johnston's account thus involves a sort of "abstracting away" from masking conditions and I believe that Smith's view was influenced by Johnston's view in this respect. However, once we move away from a single, special counterfactual to a whole range of counterfactuals as Smith does, there is (i) reason to think that we need not ignore situations with masking conditions and (ii) reason to think that we should not ignore them. To see this, consider a different case. Glass 1 is disposed to break when dropped, but it never breaks when it is dropped and the floor simultaneously turns to jelly. Glass 2 breaks in all of the same situations as glass1, but it is also disposed to break on jelly floors. Intuitively, glass 2 is more disposed to break when dropped than glass 1. And so, intuitively, how a glass performs in jelly-floor scenarios is relevant to its disposition to break: it is relevant to the degree to which an agent has the disposition in question. So there is at least some reason for thinking that we should not ignore the masking scenarios. If we hold that a glass needs to break in a suitable proportion of dropping cases, glass 1 will still have the disposition to break when dropped, even though it

Smith takes the chameleon case to show why at the outset of a C-case, where we are concerned with whether O is disposed to M when C, O may not be an intrinsic duplicate of O as it actually is. A related case suggests that at the outset of a C-case O must also differ from O as it actually is in certain extrinsic respects. Suppose that, without looking at it, Bobby would capture the chameleon just before it is viewed and would paint it blue. If we want the result that the chameleon is disposed to look green when viewed, then Smith will hold that we must abstract away from Bobby and his blue paint.

Viewing conditions relevant to the chameleon's disposition to look a certain way when viewed are thus situations in which the chameleon is a duplicate of the actual chameleon, so far as the properties of its skin are concerned, but where the chameleon lacks its peculiar psychology and is not in the presence of conditions like Bobby with his blue paint. Given that we abstract away from his psychological properties and from the presence of conditions like Bobby, the chameleon will look green in (a suitable proportion of) the relevant viewing situations.

Just as the chameleon's disposition to look a certain way when viewed is explained by his possession of certain skin properties, Smith holds that the reckless woman's ability to believe correctly while deliberating is explained by her possession of certain neural properties. Call the relevant properties *the belief forming properties*. At the outset of a deliberation case, when we are concerned with whether the woman is able to believe correctly when deliberating, we abstract away from everything except the belief forming properties. Just as we abstracted away from Bobby in the case of the chameleon, we abstract away from Black in the case of the

does not break in some of the relevant dropping cases, e.g. cases where the floor turns to jelly upon dropping. A view of dispositions is developed along these lines in Manley & Wasserman (2008).

reckless woman. Just as we abstracted away from the chameleon's peculiar psychology, we abstract away from the woman's self-hatred.⁸⁸

Smith's view, then, is that at stage 1 of responsibility assessment, i.e. in assessing whether the reckless woman was able to form the correct belief, we consider a range of deliberative situations. In all of these situations, the woman has the belief-forming properties, but she is not filled with self-hatred (and does not have other similar psychological properties that might cause her to believe incorrectly) and she is not in the presence of Black or similar conditions. In some of these situations, external circumstances differ from the actual circumstances in certain small ways. The time at which she begins deliberating might be slightly different and the way in which she asks herself the question might also be different. If, in a suitable proportion of these cases, the woman believes correctly, then she is able to believe correctly. If not, she is unable to believe correctly.

At the second stage, we consider the woman's beliefs in a similar range of deliberative situations. However, we do not "abstract away" from (at least some of) those properties that may have disrupted the woman's belief formation. Given that she is actually filled with self-hatred or that she is actually in the presence of Black with his devilish plans, we hold these facts about the woman fixed at the outset of her deliberations. If, in a suitable proportion of these situations, the woman believes incorrectly while possessing the ability to believe correctly, then we conclude

⁸⁸ I do not know how we are supposed to identify the belief forming properties and so I am unsure why *being filled with self-hatred* is not among them. The self-hatred property explains why the reckless woman forms the incorrect belief and so it is surely relevant in some way to her belief-forming mechanism, whatever it is. I will simply assume that there is some way of identifying an agent's belief forming properties in such a way that properties like being filled with self-hatred are not among them.

Here is a related point. It is not at all obvious to me that we should agree that Johnston's chameleon is disposed to look green when viewed. After all, the chameleon's psychology is relevant to the way the chameleon looks when viewed and his psychological properties are intrinsic. I am, in fact, strongly inclined to the view that the chameleon is disposed to look red when viewed. Of course, this result is unacceptable if greenness is the disposition to look green when viewed, as Smith holds: "Does the shy but powerfully intuitive chameleon lack the disposition that greenness is, that is, the disposition to look green in standard viewing conditions?" But to the extent that I'm tempted to say that the chameleon is disposed to look red when viewed, I am also tempted to say that the case is a problem for the dispositional theory of color.

that she is not blameworthy for her belief failure. If she believes correctly in a suitable proportion of these situations, then she is blameworthy.

6. Evaluating Smith's Account

I will offer two arguments against Smith's account. In this section I argue that Smith's view fails to treat like cases alike: it delivers different verdicts in cases where there is no relevant normative difference.

Smith believes that in certain versions of RECKLESS there will be no ordinary explanation for why the reckless woman failed to exercise her ability to form the correct belief. It could be that she fails to exercise this ability because she failed to vividly imagine her children's disappointment or because she fails to recall the way that she felt after making similar decisions in the past. As Smith points out however, this just pushes the question one step back. In these versions of the case, to determine whether she is blameworthy for her belief failure, we must first determine whether she was blameworthy for failing to vividly imagine her children's disappointment. So we must ask whether she had the ability to imagine her children's disappointment and whether it was failure to exercise this ability that explains her failure. At some point, Smith claims, we will either have to rest content with the claim that the reckless woman "simply blanked" or with some explanation of her failure in terms of "brain science." And Smith holds that, the availability of such a neurological explanation "give[s] us no reason whatsoever to suppose that we cannot hold the reckless woman responsible" for her belief failure. The reason, he says, is that the woman had the ability to form the belief and that her failure to exercise the ability is the relevant explanation of her belief failure.

Now suppose that the reckless woman has the ability to believe correctly. Suppose also that (i) there is no psychological explanation for her belief failure, e.g. no explanation in terms of her

self-hatred, and (ii) no explanation for her belief failure in terms of Black or similar external conditions. There is a causal explanation for her failure in terms of brain science: neuron N fires in her brain just before she begins deliberating. Suppose that it is nomically necessary that, if N fires, the woman will not form the correct belief because N fired. If the woman must vividly imagine her children's disappointment in order to believe correctly, then we can suppose that the firing of N causes the woman to fail to imagine this at every world that conforms to the actual laws.

Smith appears to hold that in the above case, where the firing of N explains the woman's belief failure, she is morally responsible. But it is hard to see that there is a relevant difference between this case and the version of the case in which the woman's belief failure is explained by her self-hatred. In both cases, there is some condition that explains her belief failure, a brain event in the first case and a state of self-hatred in the second, conditions that we can suppose the woman is unaware of and is unable to eradicate. The only relevant difference appears to concern the kinds of conditions in question: though both are intrinsic to the woman, one is a condition that we describe in psychological terms and one is a condition that we describe in neurological terms. But this difference appears rather shallow. Smith himself assumes, "for the sake of argument", that psychological states are constituted by "some state of the brain". If being in a particular psychological state is explained by being in a particular brain state (as it presumably is given Smith's assumption that brain states constitute psychological states), then, by the transitivity of explanation, being in a particular brain state explains the woman's belief failure if being in a particular psychological state does. Since there can be exculpatory explanations in terms of the woman's psychology, it thus appears that there can be exculpatory explanations in terms of the state of her brain. What justifies allowing some of these brain-level explanations,

e.g. that she is in such-and-such brain state that constitutes her self-hatred, to constitute excuses but not others, e.g. that neuron N fired in her brain?

It is unclear whether Smith would actually accept the claim that the reckless woman is blameworthy in the neural firing case. It is unclear because Smith does not provide a general recipe for answering the central question at the second stage of responsibility assessment. Presumably we must consider deliberative situations in which the woman is a psychological duplicate of the woman as she actually is, i.e. whatever psychological properties the woman actually has are properties that she must have in each deliberative situation under consideration.⁸⁹ Smith could also hold that, at the outset of a deliberation case, the woman is not only a psychological duplicate, but also a complete intrinsic duplicate of the woman as she actually is just prior to her engaging in deliberation. If so, then neuron N will fire in the woman's brain at the outset of the relevant deliberation cases. As a result, she will not believe correctly in these cases and so we should conclude that the woman is not blameworthy.

Suppose that Smith endorses the above view. While the woman is not blameworthy in the original neural firing case, there are alternative versions of the neural firing case in which the woman will be blameworthy. Suppose, for example, that neuron N does not fire prior to the deliberation or at the moment that deliberation begins. Instead, neuron N fires because of the precise way that she poses the deliberative question to herself. Had she asked the question in a different tone of voice, or with different wording, or at a slightly different time, then neuron N would not have fired. At the second stage of responsibility assessment we consider deliberation cases in which the woman is a complete intrinsic duplicate of the woman as she actually is just prior to deliberation. But in most of these cases, she poses the deliberative question to herself in

⁸⁹ "...now instead of abstracting away from her various emotions, we take her emotions into account. If, for example, she does suffer from self-hatred, does her self-hatred undermine the truth of the relevant counterfactuals?"

a slightly different way than she actually does. As a result, neuron N does not fire and she forms the correct belief. Thus, the woman is blameworthy for her belief failure.

On the resulting view, the fact that an agent's being in a particular brain state causally explains her belief failure may or may not provide the agent with an excuse. In the original neural firing case, the fact that neuron N fires in the woman's brain excuses her. But, in the second case, the fact that N fires in her brain does not excuse her. The intuitive difference between the cases is that, in the first case, the firing of N is a modally robust feature of the woman at the outset of deliberation, whereas, in the second case, the firing of N is a modally fragile feature of hers. In the first case, given that neuron N has fired in her brain before deliberating, all relevant worlds in which she deliberates about whether to have another drink are worlds in which neuron N fires. In the second case, by contrast, given that neuron N has not fired before deliberating and given that it will only fire if she poses the deliberative question *just so*, most relevant worlds in which she deliberates about whether to have another drink are worlds in which neuron N does not fire.

The resulting view is implausible. To see this, first recall that, on Smith's view, when Black causes the woman's belief failure, the agent is not morally responsible for that failure. The account delivers this consequence since in all of the relevant deliberative situations in which Black is present, he causes the woman to form the incorrect belief. But consider the following modification of the Black case: Black will cause the woman to form the incorrect belief, but only if she is in a particular fragile neural state. Suppose, for example, that Black will only cause the woman to form the incorrect belief if neuron N fires. As it happens, she poses the deliberative question to herself *just so* and neuron N fires as a result. So Black causes the woman to form the incorrect belief. Given this setup, most of the deliberative situations at the second stage will be

situations in which the woman does not pose the deliberative question to herself *just so* and so in most of the situations neuron N does not fire. Thus, since Black does not induce belief failure when neuron N does not fire, the woman will believe correctly in most of these situations. It follows that she is blameworthy for failing to form the correct belief. But this consequence is unacceptable: if the woman is not blameworthy in the original scenario involving Black, she is not blameworthy in the alternative scenario either.

7: *Objection 2 to Smith's Account*

Determining whether or not the woman has the ability to believe correctly when deliberating is a matter of determining whether or not she forms the correct belief in a certain range of deliberative situations. We focus on situations in which circumstances are not as they actually are: the temperature in the room is different or the woman poses the deliberative question to herself in a number of different ways. In this section, I want to put pressure on this feature of Smith's account. My tentative conclusion will be for the thesis *fine-graining skepticism*: if dispositional abilities are necessary for moral responsibility at all, the abilities in question must be maximally fine-grained.

It is natural to hold, in opposition to Smith, that the ability relevant to moral responsibility is not the ability to believe correctly when deliberating, but rather the ability to believe correctly when deliberating in circumstances C, where C is a maximally specific characterization of the woman's environment. Perhaps the woman would have formed the correct belief, had the temperature in the room been slightly different. But it is puzzling how the woman's performance in rooms with different temperatures is relevant to her responsibility. Why isn't the ability relevant to responsibility the ability to believe correctly when in a room of such-and-such temperature?

I will draw out this point in two ways. First, suppose that Steve wants the woman to form the incorrect normative belief. He knows that if he asks her whether she ought to have another drink in a very particular way, she will form the incorrect belief. So he asks her the question in just that particular way. She fails to form the correct normative belief. In this case, it seems that the woman is not blameworthy for her belief failure. Steve knew how to ensure that the woman believed incorrectly. Given his knowledge and motivations, the woman did not have a chance. I suggest that a natural explanation for why she is not blameworthy is that she lacks the more fine-grained ability to form the correct belief when the question is asked in just the way Steve asked it. If so, then the coarse-grained abilities identified by Smith are not the abilities relevant to moral responsibility.⁹⁰

Second, note that everyone agrees that the abilities relevant to moral responsibility must be at least relatively fine-grained. To see this, consider a few examples.

Jones is able to rescue his child when his child needs help. But his child is currently drowning and Jones cannot swim. Jones, though he has the ability to rescue his child when his child needs help, is unable to rescue his child when his child is drowning. Intuitively, Jones is excused from failing to rescue his child because he lacks this more fine-grained ability.⁹¹

Maria has the ability to sing. She sings all the time when she is in the shower, in the park, and in the car. But she blamelessly has a paralyzing variety of stage fright. So she is unable to sing on stage. Currently, Maria is on stage and something disastrous will happen if she does not

⁹⁰ Of course, Smith may be able to explain why the reckless woman is not blameworthy when Steve asks her the question in just the way that he does. On his account, the woman has the ability to form the correct belief when she is asked the deliberative question. But she does not pass the second stage of responsibility assessment. Given the presence of Steve, she will fail in most deliberative contexts since Steve will ask her the question *just so*. The point is not that Smith does not have the resources to classify the case correctly. Rather, the case is supposed to directly elicit the intuition that it is the more fine-grained ability that matters, the intuition that her lack of the fine-grained ability is excusing.

⁹¹ I am supposing that Jones is blameless for being unable to swim and for being in a situation in which his child is drowning.

sing. She is blameless for being in these circumstances. She does not sing and disaster strikes. Intuitively, while Maria has the coarse-grained ability to sing, she is excused because she lacks the more fine-grained ability to sing on stage.

Once we make the move to *relatively* fine-grained abilities, the move to *maximally* fine-grained abilities begins to appear irresistible. Consider Maria's cousin, Martha. Martha has the ability to sing and the ability to sing on stage. But she is unable to sing when on stage and she has not slept enough (or to sing when her parents are in the audience and the light is shining in her eyes in just this particular way). If she is blameless for not having slept enough (or for her parents' being in the audience and for the light's shining *just so*), then intuitively she is excused for failing to sing. At least, she is excused if Maria is. And once we have gone this far, it is unclear how we are to draw a principled distinction between the ability to sing on stage after sleeping for only four hours, with ones' parents in the audience, etc. and the ability to sing on stage when C, where C is a maximally specific characterization of the agent's circumstances.

Of course, in ordinary contexts in which an agent's moral responsibility is in question, we typically concern ourselves with whether or not the agent possesses an ability that is characterized in a relatively fine-grained way, rather than an ability that is maximally fine-grained. This is because we are sometimes actually in a position to know that a given agent possesses or lacks a relatively fine-grained ability, like the ability to sing on stage. We know that some people suffer from terrible stage fright and so lack this ability. But we are presumably never in a position to know whether or not an agent possesses a maximally fine-grained ability. While this epistemic distinction between propositions attributing relatively fine-grained abilities to agents and propositions attributing maximally fine-grained abilities to agents might help to explain our actual practices, it is not plausibly normatively significant.

8. *The Way Forward*

I do not take the argument for *fine-graining skepticism*, the thesis that the dispositional abilities necessary for moral responsibility are maximally fine-grained, to constitute a completely devastating objection to the dispositional compatibilist position. Instead, I see the argument as a challenge that the dispositional theorist must address. Extant dispositional theorists have held, not only that normatively significant abilities are dispositions, but also that normatively significant abilities are those that we typically have in mind when assessing an agent's blameworthiness, those abilities that we typically express with ability sentences in moral responsibility contexts. Fine-graining skepticism puts pressure on this feature of the traditional dispositional compatibilist position.

The dispositional theorist may attempt to answer the fine-graining challenge in the following way. If we begin by focusing on a case in which an agent who acts wrongly possesses a range of coarse-grained abilities to do what he should, but who lacks (something close to) the maximally fine-grained ability to do what he should, then we will form a firm judgment that the agent in question deserves harm. Moreover, the dispositional theorist will perhaps insist that this judgment enjoys the status of a Moorean fact, a fixed data point in terms of which the theory of moral responsibility is constrained. The fine-graining skeptic will, of course, be unimpressed with this reply. He takes the fine-graining challenge to show that our ordinary reaction to (certain cases of) wrongdoing is unjustified. The challenge itself casts serious doubt on the dispositional theorist's alleged Moorean fact.

It may be that the way forward is to first take a step back. The compatibilism dispute has been reduced to a normative dispute over the conditions under which an agent deserves harm. And to this point, I have simply relied upon our pre-theoretic grasp of the notion *S deserves harm for Xing*. Perhaps with an informative account of the concept of deserving harm on the table we

will be in a position to identify which party to the dispute, the traditional dispositional theorist or the fine-graining skeptic, is making a mistake.

It is certainly worth examining the compatibilism dispute through the lens of a more sophisticated account of the ethics of deserving harm. I am somewhat skeptical, however, that progress lies in this direction. For in developing the complete theory of the ethics of deserving harm, we are likely to find ourselves in the same dialectical standoff. Elucidating the notion of an agent's deserving harm for an action will largely be a matter of respecting our judgments about particular cases in which an agent does or does not deserve harm. We should expect that some theorists, those inclined to the traditional dispositional theory, will be unwilling to give up their judgment that an agent's lacking extremely fine-grained abilities does not show that the agent does not deserve harm. And we should expect that others, those inclined to fine-graining skepticism, (i) will reject the dispositional theorist's judgment and (ii) will be unwilling to give up the claim that, whenever an agent is excused in virtue of lacking a relatively fine-grained ability, a similar agent should be excused in virtue of lacking an ability that is slightly more fine-grained.

I hope that my skepticism about our ability to make progress on the compatibilism dispute by first illuminating the concept of deserving harm is unjustified. For if it is justified, then it is tempting to think that there is no way forward – there are just two ways of reaching a state of equilibrium, with nothing further to be said.

Bibliography

- Arpaly, Nomy (2003). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford University Press.
- Austin, J.L. *Ifs and Cans* (1956). British Academy Lecture. Oxford University Press.
- Balaguer, Mark (2010). *Free Will as an Open Scientific Problem*. MIT Press.
- Beebe, Helen and Mele, Alfred (2002). "Humean Compatibilism," *Mind* 111.
- Brueckner, Anthony (2008). "Retooling the Consequence Argument," *Analysis* 68.
- Campbell, Joseph (2007). "Free Will and the Necessity of the Past," *Analysis* 67.
- Campbell, Joseph (2008). "Reply to Brueckner," *Analysis* 68.
- Campbell, Joseph (forthcoming). "Free Will and the Consequence Argument," in *Internet Encyclopedia of Philosophy*.
- Chalmers, David (2011). "Verbal Disputes," in *Philosophical Review* 120.
- Clarke, Randolph (2009). "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism," in *Mind*.
- Edgington, Dorothy (1995). "On Conditionals," in *Mind*.
- Fara, Michael (2009). "Masked Abilities and Compatibilism," in *Mind*.
- Fine, Kit (2001). "The Question of Realism," *Philosopher's Imprint* 1.
- Fischer, John (1994). *The Metaphysics of Free Will*. Oxford: Blackwell.
- Fischer, John, Kane, Robert, Pereboom, Derk, and Vargas, Manuel (2007). *Four Views on Free Will*. Oxford: Blackwell.
- Ginet, Carl (1990). *On Action*. Cambridge University Press.
- Huemer, Michael (2000). "Van Inwagen's Consequence Argument," in *Philosophical Review* 109.
- Huemer, Michael (2004). "Elusive Freedom? A Reply to Helen Beebe," in *Philosophical Review* 113.
- Hume, David (1739/2000). *A Treatise of Human Nature*. Oxford University Press.
- Hofweber, Thomas (2009). "Ambitious, yet Modest, Metaphysics," in *Metametaphysics*:

- New Essays on the Foundations of Ontology* (eds. Chalmers, Manley, and Wasserman).
- Kane, Robert (1996). *The Significance of Free Will*. Oxford University Press.
- Kment, Boris (2006). "Counterfactuals and Explanation," in *Mind* 115,
- Kratzer, Angelika (1977). What "must" and "can" must and can mean. *Linguistics and Philosophy* 1.
- Lehrer, Keith (1968). "Cans without Ifs," in *Analysis*.
- Lewis, David (1976). "The Paradoxes of Time Travel," in *American Philosophical Quarterly*.
- Lewis, David (1979). "Counterfactual Dependence and Time's Arrow," in *Nous* 13.
- Lewis, David (1981). "Are We Free to Break the Laws?," in *Theoria* 47.
- Manley, David & Wasserman, Ryan (2008). "On Linking Dispositions with Coniditonals," in *Mind* 117.
- Martin, C.B. (1994). "Dispositions and Conditionals," in *Philosophical Quarterly* 44.
- McKay, Thomas and Johnson, David (1996). "A Reconsideration of an Argument against Compatibilism", in *Philosophical Topics* 24.
- Mele, Alfred. "Agents' Abilities," in *Nous* 37.
- Merricks, Trenton (2008). "Replies to Cameron, Schaffer, and Soames," in *Philosophical Books* 49.
- Moore, G.E. (1912/2005). *Ethics*. Oxford University Press.
- O'Connor, Timothy (1993). "On the Transfer of Necessity," in *Nous* 27.
- Rosen, Gideon (2010). "Metaphysical Dependence: Grounding and Reduction," in Bob Hale and Aviv Hoffmann (eds.), *Modality: Metaphysics, Logic, and Epistemology*, Oxford University Press.
- Rosen, Gideon (unpublished). "The Alethic Conception of Culpability."
- Scanlon, Thomas (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Harvard University Press.
- Schaffer, Jonathan (2004). "Counterfactuals, Causal Independence and Conceptual Clarity," in *Analysis*.

- Schnieder, Benjamin (2004). "Compatibilism and the Notion of Rendering Something False," in *Philosophical Studies* 117.
- Smart, J.J.C. (1963). "Free Will, Praise, and Blame," in *Mind* 70.
- Smith, Michael (2003). "Rational Capacities, Or: How to Distinguish Recklessness, Weakness, and Compulsion," in Sarah Stroud & Christine Tappolet (eds.), *Weakness of Will and Practical Irrationality*. Clarendon Press.
- Soames, Scott (2008). "Truthmakers," *Philosophical Books* 49.
- Spencer, Jack (unpublished). "Able to Do the Impossible."
- Strawson, P.F. (1962). "Freedom and Resentment," in *Proceedings of the British Academy*.
- Van Fraassen, Bas (1989). *Laws and Symmetry* (Oxford University Press).
- Van Inwagen, Peter (1983). *An Essay on Free Will* (Oxford University Press).
- Van Inwagen, Peter (2000). "Free Will Remains a Mystery," in *Philosophical Perspectives* 14.
- Van Inwagen, Peter (2008). "How to Think about the Problem of Free Will," in *Journal of Ethics* 12.
- Vihvelin, Kadri (1988). "The Modal Argument for Incompatibilism," in *Philosophical Studies* 53.
- Vihvelin, Kadri (2004). "Free Will Demystified: A Dispositional Account," in *Philosophical Topics*.
- Vihvelin, Kadri (2008). "Compatibilism, Incompatibilism, and Impossibilism," in Theodore Sider, John Hawthorne, and Dean W. Zimmerman (eds.), *Contemporary Debates in Metaphysics*. Blackwell.
- Vihvelin, Kadri (2009). "How to Think about the Free Will/Determinism Problem," available at Vihvelin.com.
- Vihvelin, Kadri (2011). "Arguments for Incompatibilism", in *Stanford Encyclopedia of Philosophy*.
- Wallace, Jay (1998). *Responsibility and the Moral Sentiments*. Harvard University Press.
- Warfield, Ted (2003). "Compatibilism and Incompatibilism: Some Arguments" in M.J. Loux and D.W. Zimmerman, eds., the *Oxford Handbook of Metaphysics*. NY: Oxford UP.
- Widerker, David (1987). "On An Argument for Incompatibilism," *Analysis* 47.