

NON-REDUCTIVE PHYSICALISM AND FREE WILL

By

BENJAMIN WHITTINGTON

A thesis submitted to the University of Birmingham for the degree of

Doctor of Philosophy

Department of Philosophy

School of Philosophy, Theology and Religion

College of Arts and Law

The University of Birmingham

November 1st, 2023

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

ABSTRACT:

Imagine that God has two shelves of ingredients. With these two shelves, God creates the whole cosmos. The first shelf exclusively contains physical ingredients (e.g., electrons). The second shelf is a shelf of non-physical ingredients (e.g., ghostly ectoplasm). The central question of this thesis is whether or not God needs to reach up to the second shelf (the non-physical shelf) in order to create free creatures. Put another way, the central question of this thesis is whether physicalism- particularly, non-reductive physicalism- is sufficient for free will (or free agency).

To address the above question, I develop a novel version of the modal consequence argument purporting to establish that non-reductive physicalism and free will are incompatible. This argument consists of three crucial elements: (i) Rule Beta, according to which if you cannot change some state of affairs x and x necessarily entails y , then, therefore, you cannot change y , (ii) the thesis that, given non-reductive physicalism, there is nothing anyone can do that would change the microphysical states on which a person's decisions depend on, (iii) the thesis that if a person cannot do anything other than what she actually does, then she cannot have free will.

To argue for element (i), I utilize Alexander Pruss's argument from weakening to show that Rule Beta is a valid modal inference rule and defend it from an objection from Michael Mckenna and Derk Pereboom. I defend element (ii), which functions as the fixity principle, by building on L.A. Paul's argument from additivity and defending it from objections anticipated by Christian List and Jessica Wilson. Element (iii) implies that free will necessarily requires alternative possibilities. To establish this premise, I utilize and augment John Maier's modal theory of abilities to craft a new definition of alternative possibilities and defend it from Frankfurt-style Objections. I conclude that the modal consequence argument gives us a reason to think that God does indeed need to reach up to the second shelf.

ACKNOWLEDGEMENTS:

Looking back over the time my project took, there are several people I am grateful toward. I would like to thank my supervisors, Yujin Nagasawa and Jussi Suikkanen. I will forever appreciate your patience, advice, and encouragement for me during this project.

I would also like to thank my wife Emily for encouraging and being with me from start to finish. Getting a PhD is an adventure, not only for the student, but for the student's family. So, I am grateful for your patience and encouragement as well.

I am also thankful for various classmates, friends, correspondents, and conference presentation attendees who helped me on this journey. In particular, I am grateful toward Eric Yang, Roger Turner, Randy Everist, Taylor Cyr, and Nathan Dowell.

INSCRIPTION:

It is as impossible for the subtlest philosophy as for the commonest reasoning to argue freedom away. Philosophy must therefore assume that no true contradiction will be found between freedom and natural necessity in the same human actions, for it cannot give up the idea of nature any more than that of freedom. Nevertheless, even though we should never be able to comprehend how freedom is possible, we must at least remove this apparent contradiction in a convincing manner. For if the thought of freedom contradicts either itself or nature, which is equally necessary, it must in competition with physical necessity be entirely given up. -Immanuel Kant

TABLE OF CONTENTS

INTRODUCTION.....	8
CHAPTER ONE. DEFINING FREE WILL AND PHYSICALISM.....	14
Introduction	14
1.1 Free Will	14
1.1.1 Alternative Possibilities and Conscious Quausation.....	20
1.1.2 Formal Definition of Free Will	26
1.2 Determinism.....	28
1.3 Physicalism	31
1.3.1 Global Supervenience, Fundamental Physicality, and Nomological Closure	32
1.3.2 Mental Ontological and Causal Irreducibility	39
Conclusion	58
CHAPTER TWO. THE P-MODAL CONSEQUENCE ARGUMENT.....	59
Introduction	59
2.1 The Consequence Argument	60
2.2 The Modal Consequence Argument	63
2.2.1 Agglomeration	68
2.3 Pruss' Modal Consequence Argument.....	72
2.3.1 Beta-Best.....	72
2.3.2 The P-Modal Consequence Argument.....	78
2.4 Mckenna and Pereboom's Objection	80
2.4.3 The Formal Argument Again.....	85
Conclusion	87
CHAPTER THREE. THE B-MODAL CONSEQUENCE ARGUMENT.....	89
Introduction	89
3.1 The Relevance of Physicalism	90
3.2 The B-Modal Consequence Argument	92
3.2.1 Principles Common To Both (The P and B) Modal Consequence Arguments	93
3.2.2 Principles Only In The B-Modal Consequence Argument	95
3.2.3 The Formal Argument.....	100
Conclusion	103
CHAPTER FOUR. THE FIXITY OF THE MICROPHYSICAL CONFIGURATION.....	104
Introduction	104
4.1 The Fixity of the Microphysical Configuration	104
4.2 Christian List's Intra-Level Mental Causation	114
4.3 Jessica Wilson's Subset View of Mental Causation	120
4.4 The Problem of Additivity Again	127
4.4.1 A General Problem with Subset Views	127
4.4.2 The Problem.....	128
Conclusion	133
CHAPTER FIVE. ALTERNATIVE POSSIBILITIES.....	134
Introduction	134
5.1 John Maier's Modal Theory of Abilities	136
5.1.1 Modal Abilities	137
5.1.2 Options and Abilities	138

5.1.3	The Robustness Objection	144
5.1.4	Maier and the Robustness Objection	147
5.1.5	Maier on the Consequence Argument.....	149
5.1.6	My Response to Maier	156
5.1.7	Are Options Primitive?	160
5.2	The Frankfurt-styled Objection to The Consequence Argument.....	166
5.2.1	The Ordinary Frankfurt-styled Argument.....	168
5.2.2	Helen Steward and Maria Alvarez's Response.....	172
5.2.3	The Standard Frankfurt-styled Argument	174
5.2.4	The Dilemma Defense	178
5.2.5	Mele and Robb's Frankfurt Case	182
5.2.6	Two Responses	184
5.2.7	John Martin Fischer's New Frankfurt Argument.....	188
5.2.8	My Response.....	194
	Conclusion	200
CHAPTER SIX: INTERACTIONIST DUALISM AND FREE WILL, FREE WILL		
IMPOSSIBILISM, OR FREE WILL REVISIONISM		201
	Introduction	201
6.1	Review	202
6.2	Interactionist Dualism	205
6.2.1	The Argument from Conservation Laws	208
6.2.2	Three Dualistic Responses.....	211
6.2.3	The Dualistic Ingredient for Free Will.....	219
6.3	Free Will Impossibilism.....	221
6.3.1	Moral Responsibility	222
6.3.2	Meaning of Life	223
6.3.2	Rationality	225
6.4	Revisionism.....	226
	Conclusion	229
REFERENCES.....		230

INTRODUCTION

Imagine that God has two shelves of ingredients. With these two shelves, God creates the whole cosmos. The first shelf exclusively contains physical ingredients (e.g., electrons, quantum fields, and atomic nuclei). The second shelf contains non-physical ingredients (e.g., ghostly ectoplasm, souls, and proto-consciousness). The central question of this thesis is whether or not God would need to reach up to the second shelf (i.e., the non-physical shelf) in order to create free creatures. Put another way, the central question of this thesis is whether physicalism- particularly, non-reductive physicalism- is sufficient for free will (or free agency).

To address the above question, I develop a novel version of the modal consequence argument purporting to establish that non-reductive physicalism and free will are incompatible. This argument consists of three crucial elements: (i) Rule Beta, according to which if you cannot change some state of affairs x and x necessarily entails y , then, therefore, you cannot change y , (ii) the thesis that, given non-reductive physicalism, there is nothing anyone can do that would change the microphysical states on which a person's decisions depend on, (iii) the thesis that if a person cannot do anything other than what she actually does, then she cannot have free will.

To argue for element (i), I utilize Alexander Pruss's argument from weakening to show that Rule Beta is a valid modal inference rule and defend it from an objection from Michael Mckenna and Derk Pereboom. I defend element (ii), which functions as the fixity principle, by building on L.A. Paul's argument from additivity and defending it from objections anticipated by Christian List and Jessica Wilson. Element (iii) implies that free will necessarily requires alternative possibilities. To establish this premise, I utilize and augment John Maier's modal theory of abilities to craft a new definition of alternative possibilities and defend it from Frankfurt-style Objections. I conclude that

the modal consequence argument gives us a reason to think that God does indeed need to reach up to the second shelf.

To be clear, when I discuss the “physical ingredients” of agency I am referring to ingredients compatible with physicalism. Physicalism is roughly the view, as David Lewis and Rae Langton say “that physics—something not too different from present-day physics, though presumably somewhat improved—is a comprehensive theory of the world, complete as well as correct. The world is as physics says it is, and there’s no more to say” (Lewis & Langton, 1998, pp. 333-334). Physicalism is a relatively straightforward metaphysical thesis, but there are different versions of the view under the broad umbrella of physicalism. The majority of this thesis will focus on the incompatibility of what is often called Type-B physicalism, or *non-reductive physicalism* (or, as some more recent philosophers have referred to it, “weak emergentism”), and free will.

This thesis will seek to prove the following claim: Necessarily, if non-reductive physicalism is true, then there cannot exist creatures with free will. In other words (using possible world language), there is no possible world where non-reductive physicalism is true and persons with free will exist. From this, I will consider three available options regarding free will: Either (I) we have free will and dualism is true, (II) free will is impossible in the actual world (because non-reductive physicalism is true), or (III) we should be revisionists about free will.

Part one of this thesis will aim to formulate the incompatibility argument that I shall defend in the remainder of this thesis. **Chapter One** will briefly provide some background to the thesis by defining three key terms: free will, determinism, and physicalism. I begin by discussing two key conditions for free will. The first is that in order for a person P to freely will some outcome ϕ at t , P must have been able to decide $\neg \phi$ at t . The second is that in order for P to freely will some

outcome ϕ at t , ϕ must have been caused by P's conscious intention. After discussing these two conditions, I will move to define the concept of determinism, which can be roughly stated as the thesis that, given any particular completed state of the world (at a given time t) in conjunction with the laws of nature, there is only one modally possible future and past. Finally, I will work to define the theory of physicalism. I will discuss two key doctrines for physicalism: Global Supervenience and Nomological Closure. I will also distinguish, using David Chalmers' terminology (1996, 2010), Type-A versions of physicalism from Type-B versions of physicalism in order to argue that the best candidate for a version of physicalism that is compatible with free will is one that accepts a doctrine called Mental Ontological and Causal Irreducibility, which is an essential feature of non-reductive physicalism.

In **Chapter Two**, I will begin to consider an argument that is a good candidate to test the compatibility of non-reductive physicalism and free will: the consequence argument. Chapter Two's ultimate goal is to explore a modally valid way to present the traditional consequence argument (against the compatibility of determinism and free will). I begin by introducing the original consequence argument (Van Inwagen, 1983), and some of the more basic intuitions undergirding the argument (Huemer, 2021). Then, I will introduce the more specific modal versions of the consequence argument, and the chief objection to the modal consequence argument that all versions will have to address: the problem of agglomeration (McKay & Johnson, 1996). I will additionally explain one contemporary version of the modal consequence argument – which I refer to as the P-MODAL CONSEQUENCE ARGUMENT – from Alexander Pruss (Pruss, 2013), which I argue is able to overcome the problem of agglomeration. Finally, I will discuss an objection to the validity of the P-MODAL CONSEQUENCE ARGUMENT from Michael Mckenna and Derk Pereboom (2016), and present a response

Having shown that the P-MODAL CONSEQUENCE ARGUMENT is valid, in **Chapter Three** I will modify the argument in certain ways in order to make the argument about the compatibility of non-reductive physicalism and free will, which I will refer to this argument as the B-MODAL CONSEQUENCE ARGUMENT. This new argument will roughly state that because (I) non-reductive physicalism entails that a microphysical state S necessitates a person P's intentional decision in a choice situation, and (II) no one has the ability to intentionally make S not obtain (because microphysics is nomologically closed), then, therefore (III) necessarily, no one can have free will.

Part two of the thesis will then defend the controversial aspects of the B-MODAL CONSEQUENCE ARGUMENT. There are two controversial premises, and I will discuss each in separate chapters.

In **Chapter Four**, I will defend the second premise of the argument, which I call the fixity of the microphysical configuration principle, which is that (necessarily, assuming non-reductive physicalism) no one has the ability to intentionally make S not obtain (where S is the microphysical state necessitating a person's intentional decision in a choice situation).¹ I will show that no one can mentally or consciously make a microphysical state not obtain (or an event not to occur) given non-reductive physicalism. I will, however, also explain that this argument does not claim that mental causation is impossible. I will explain two non-reductive physicalist views of mental causation. The first is from Christian List and Peter Menzies (2016; List, 2019), who argue that causation always occurs intra-level (and never inter-level). This, I show, is a good response to Jaegwon Kim, but not a good objection to the argument for the Fixity of the Microphysical

¹ We can also say no one can change S.

Configuration. The second is from Jessica Wilson (2020) who has developed a subset model of mental causation where mental states (or events, or properties, or processes) have a distinct subset of causal powers from their microphysical realizer state (while the individual constituents of those sets are all shared). This model of mental causation, I explain, is an effective response to the argument for the Fixity of the Microphysical Configuration. However, I argue that this model is implausible because of its reliance on counterfactual causation which I argue (via what L.A. Paul calls the Problem of Additivity) is not plausible for mental causation in particular. By the end of this chapter, I will have proven that the key premise (and probably the most controversial aspect) of the B-MODAL CONSEQUENCE ARGUMENT is true.

Next, I move to the second key controversial element of the B-MODAL CONSEQUENCE ARGUMENT, which is the implicit assumption that free will necessarily requires alternative possibilities. This is an aspect that, I believe, is essential to all versions of the modal consequence argument, but is seldom defended or acknowledged explicitly. In **Chapter Five** I will defend this premise and try to show that it is a sensible condition of free will. To do this, I utilize John Maier's (2015; 2020; 2022) modal theory of abilities and explain how alternative possibilities could plausibly matter for one's agency as a whole. In this section, I will outline one criticism of the consequence argument from Maier (2022) and offer a response. However, my main purpose here is to clarify the condition of needing alternative possibilities for free will. After this, I will explain a prominent objection to this condition from the success of so-called Frankfurt-cases.

In Chapters Four and Five, I will have, thus, defended the two main controversial elements of the B-MODAL CONSEQUENCE ARGUMENT. Finally, in **Chapter Six**, I consider what we should do next. If we accept the argument, then there are three clear available options of how we should think about free will in the actual world: the conjunction of dualism and free will, non-reductive

physicalism and free will impossibilism, or a kind of non-traditional revisionist perspective on free will.

Before I discuss the B-Modal Consequence Argument, I need to define what I mean by free will, determinism, and physicalism, which is where I will turn first in this thesis.

CHAPTER ONE. DEFINING FREE WILL AND PHYSICALISM

Introduction

I am going to clarify three key concepts in this chapter that will be used in this thesis: free will, determinism, and physicalism. This will help set up the subsequent chapters and discussions. I am going to begin by defining free will (§1.1), and then briefly discuss determinism (§1.2). After this, I will define the notion of physicalism (§1.3).

1.1 Free Will

To begin, I take free will to primarily concern mental actions. For example, I take free will to be concerning phenomena like Ben choosing to attend the University of Birmingham (as opposed to Durham). I will start by discussing the notion of action generally in order to make sure the terminology and concepts are clear, and then move to the notion of free will.

I will begin my discussion of actions with an example borrowed from Paul Pietroski (2000 , pp. 1-2): on April 14th 1865 the American actor (and confederate sympathizer) John Wilkes Booth assassinated the sitting United States President (i.e., Abraham Lincoln) at the Ford Theatre. We can then consider the following four statements:

- (a) John Wilkes Booth pulled the trigger on April 14th 1865.
- (b) John Wilkes Booth shot Abraham Lincoln on April 14th 1865.
- (c) John Wilkes Booth killed Abraham Lincoln on April 14th 1865.
- (d) John Wilkes Booth decided to kill Abraham Lincoln on April 14th 1865.

The first three statements (a-c) are events concerning John Wilkes Booth's physical or bodily movements (and the results of them), whereas the fourth (d) concerns his conscious decision on the day he murdered Lincoln.² Of course, (a)-(c) could have still been true if (d) were false (e.g., if Booth accidentally or mistakenly shot Lincoln), and (d) could have still been true if (a)-(c) were false (e.g., if Booth's pistol was missing its trigger and thus was broken on April 14th 1865, so that when Booth pulled it out to shoot Lincoln, he merely looked foolish).

Although there are differences between (a)-(d), they are all representing Booth's intentional actions. Roughly speaking, actions are events that are appropriately caused by mental events (or other causal *relata* connected to mentality, such as mental states or powers).³ It is important for any definition of action to include its connection to mentality or consciousness. For example, consider a bodily movement that is intuitively not an action, like a nervous tick or a muscle spasm. Muscle spasms and nervous ticks are not intentional actions, they are merely bodily movements. What makes an intentional action distinct from these mere bodily movements is the agent's mental event or state's relevance in the intentional action. For example, we usually associate Booth's

² I imagine the Booth made his plan before April 14th, and thus made a decision prior to that day, however he still made a decision *on that day*.

³ Free will theorists (and metaphysicians generally) disagree on the *relata* of causation, thus this definition of action needs to be broadly construed. I do not think this conception of action changes much based on which theory of causation is correct, the basic idea stands. However, the metaphysical definition of action will change based on one's theory of causation (and, hence, mental causation). For example, E.J. Lowe (2005) denies that only events can be causal *relata* and argues facts or states can be *relata* as well. Helen Steward (2012) argues that actions are not events or states, but processes (which are a kind of causal *relata*). Numerous recent authors have picked up on a causal power ontology in place for a more Humean or event based causal ontology, for an overview of this position see Stephen Mumford (2010).

mental state (e.g., his desire to kill Lincoln and his belief that shooting Lincoln would kill him) with his shooting Lincoln to conclude that he *killed* Lincoln. His body did not merely move.

There is one possible point of confusion that we need to be clear on. Sometimes philosophers think of actions as exclusively bodily movements (e.g., Booth pulling the trigger), but this is not right (or, at least, this is not the way I am thinking of the term “action”). For example, Booth *planned* and *conspired* with his co-conspirators to kill Lincoln before and on April 14th. While it is true that there are numerous bodily movements associated with planning and conspiring (e.g., rubbing one’s chin excessively), the core element in planning and conspiring are Booth’s mental events and states (i.e., his planning and deciding). Booth’s planning and deciding to kill Lincoln should be considered an action (or actions) even though it is not necessarily involving bodily movements.

We would hold Booth responsible for these plans and decisions independently of his physical movements. For example, American law criminalizes *conspiracy* irrespective of any physical movements. Imagine that Booth had a change of heart right before he pulled the trigger. He would certainly feel great remorse and guilt *for his plans and decisions*. Our decisions and plans are the things we control the most. To change the illustration briefly, when I *try* to score a goal playing hockey, I have a good deal of control over the puck-movement and such, but the aspect I have the most control over is whether or not I am *trying* to score (my mental state).

The core of an action – whether a mental action (e.g., Booth deciding to kill Lincoln) or a physical action (e.g., Booth pulling the trigger) – is that it necessarily has a mental component. In presenting

action in this manner, I am presenting what many contemporary philosophers call a *causalist* view of action, which understands actions in terms of causal concepts.⁴

One of the most influential arguments for *causalism* -often referred to as Davidson's Challenge (1963) - can be illustrated with Lilian O'Brien's example of an agent with multiple desires. Here is O'Brien's (2014, p. 13) example:

Jane is walking on the street with her boss. Jane desires to impress her and believes that giving some money to a collector for UNICEF is a good means to that end. But she feels ashamed of this self-centered desire and rejects it as an unworthy basis for action. Then she notices a UNICEF poster depicting children who look sad and hungry. Jane is struck by their plight and desires to help them. She acts on this desire. Jane's boss is impressed.

In this example, Jane has two important desires, the desire to impress her boss and the desire to help children in need. Either desire appears adequate to explain her action in this scenario. However, only one of these desires actually does explain her action. Davidson Challenge is simply this: what is the difference between claiming that 'Jane A'ed because she wanted to help children in need' and 'Jane A'ed because she wanted to impress her boss'? In other words, what makes it the case that the former is true, and the latter is false? Davidson argued that the most natural way

⁴ Prior to Davidson's famous essay "Actions, Reasons, and Causes" (1963), most philosophers (i.e., mainly the Wittgensteinians) believed that what made actions distinct from other events was merely the explanatory role, which mental events and states serve in explaining actions by rationalizing them as reasons (e.g., I went to the store because I wanted milk). However, prior to Davidson, these reasons were presumed to be fundamentally non-causal explainers of actions (Taylor, 1964). Precisely what a non-causal explainer (a "non-causal bringing about," as some said) in this context is not always clear. However, some philosophers still argue that reason explanations are not reducible to causal relationships (McLaughlin, 2013).

to respond to this question is by saying that Jane's desire to help needy children *caused* her action in some sense. O'Brien summarizes the causalist position:

One key to the mystery of intentional action is control – intentional action involves control by the agent over her body and over the attainment of a goal. A popular – some might say wildly popular – way of characterizing such control is in terms of mental events causing bodily movements... For causalists, the difference between a tic and an intentional waving lies in its causal history: the latter, but not the former, is caused by an intention to wave or a desire to wave. As Alfred Mele (1997) puts it, just as a burn is not a sunburn unless it was caused by the sun, a movement is not an intentional action unless it was caused by a desire or intention. (O'Brien, 2014, 12)

The causalists' reply to the Davidson Challenge thus seems highly intuitive, and therefore I consider it a good starting point for constructing a metaphysical way of understanding free will.

There are two further specifications I need to make about actions. Firstly, one common causalist view is that actions are events (whether physical or mental) caused appropriately by mental events. This is how I have been primarily discussing actions thus far. This conception is put forward by theorists like Donald Davidson (1963) and Al Mele (1992; 2003). To put the John Wilkes Booth example earlier in this schema, John Wilkes Booth desires to kill Lincoln and believes that shooting him (via pulling the trigger) would kill Lincoln; As a result, John Wilkes Booth kills Lincoln. This action fits well with this first causalist picture of action.

However, some causalists take bodily movements to be *effects* of actions. On this view, actions are exclusively a particular kind of mental phenomena (whether an event or something else). The

causalists who defend this view include E.J. Lowe (2008) and Paul Pietroski (2000). Pietroski explains this view with the following example:

In my view, many bodily motions are already effects of actions. Shifting to a non-violent example, suppose Nora reached for her umbrella because she came to believe that it was raining. I think the acquisition of her belief caused an action/trying that caused a bodily motion, which caused the motion of the umbrella. Nonetheless, mental causes of actions can still be distal causes of bodily motions, which have non-mental effects (2000, 5).

On this view, actions are causal processes or causal *relata* themselves (possibly because they are specific kinds of powers in powerist views of causation). To put the John Wilkes Booth example earlier in this schema, John Wilkes Booth's *decision* to kill Lincoln on April 14th *caused* his pulling the trigger (and subsequent events), which caused Lincoln's death, and it is Booth's decision that is really his action.

These are thus two different ways of theorizing about the metaphysics of actions within the causalist framework. For the purposes of this thesis, it does not particularly matter which is right. According to both views, particular mental features (i.e., mental states and events) are essential to every intentional action (e.g., Booth's pulling the trigger or his decision to kill Lincoln on April 14th).

To summarize, so far I have characterized an action as being inherently intentional. That is, an action is something that essentially has a mental state or event as part of its description. Without the mental state or event, it is not an action. This is true whether or not we focus on a mental action (e.g., Booth's planning to kill Lincoln) or a physical action (e.g., Booth's pulling the trigger). I

have also framed actions as inherently causal. So, when I choose to go to the refrigerator to get some cheese, my intention *causes* my bodily movements and such.

1.1.1 Alternative Possibilities and Conscious Quasation

In thinking about free will specifically now, I will start by assuming that mental actions (e.g., Booth's decision to kill Lincoln) are the paradigmatic things to be explained rather than physical actions (e.g., Booth's pulling the trigger). In other words, I take free will concern primarily *mental actions* primarily, and then, as an extension, actions more broadly speaking.⁵

When we think about mental actions, it is natural to suggest that free will concerns *willing* or *intending* some particular outcome to occur in a particular choice situation. (Willing is an old-fashioned term; some useful modern synonyms would be *deciding*, *intending*, or *choosing*). For example, I can decide to attend the University of Birmingham or intend to marry Emily on December 13th (both very important choices), or, as previously mentioned, John Wilkes Booth can decide to kill Abraham Lincoln on April 14th 1865. These would be examples of intending or deciding particular outcomes in choice situations.⁶

There are two senses in which we intend an outcome in choice situations. Firstly, sometimes we intend outcomes in the future. For example, I might form an intention to go to the gym more

⁵ Clarke (2003, 121) calls proponents of prioritizing mental actions as *willists* as opposed to the *actionists*, who take physical actions to be the paradigmatic features of freedom. In this terminology, I do assume a *willist* account of free will.

⁶ Additionally, there are hypothetical scenarios where I could intend some outcome, but not do so freely. For example, if Emily hires a hypnotist to manipulate me into asking her to marry me on December 13th (via hypnotic suggestion), this would not be a free willing.

frequently next year. Secondly, we also form intentions by making decisions about what is right in front of us in the present (or near-present). For example, John Wilkes Booth chose to kill Abraham Lincoln *on* April 14th 1865. Planning for the future is related to free will, but is notably different than the planning or forming of intentions we make in the present. I take cases where we make decisions about what is right in front of us (e.g., Booth's choice) as paradigmatic free will cases.

Additionally, we describe decisions as freely willed only in choice situations where we believe we have more than one option we can choose right in front of us (in the present). So, when Emily chooses to attend Birmingham (as opposed to Durham or Cambridge) for university, she intuitively believes she made a free decision. When Emily chooses to hand over her wallet to a mugger who threatens to kill her, she does not intuitively think of her decision as freely willed. John Martin Fischer (1994, p. 190) likens our intuitive conception of free will (i.e., the "folk conception") to a decision on a branching pathway:

We naturally think of the future as open. We think of the future as containing various paths that branch off one past; although we know we will travel along just one of these paths, we take it that some of the other paths are (at least sometimes) genuinely accessible to us. In deliberating and deciding on a course of action, we intuitively think of ourselves (at least sometimes) as determining which path to take, among various paths we could take.

Here, the branching pathway is a metaphor for the choice situations we, as agents, are in. Our pre-theoretical belief about free will is that, in choice situations, we take it for granted that at least sometimes:

- (I) Both options in front of us are genuinely accessible, [and]
- (II) it is our decision (or intention) that determines which outcome obtains or occurs.

I need to discuss the ideas expressed in (I) and (II) more in order to formulate a metaphysical definition of free will out of these intuitive concepts.

When we think about (I), we intuitively believe that the future is open in the sense that the options we consider are genuine possibilities.⁷ Having genuine open options available is important to free will because we seem to care about what we could have done in the past and what we can do in the future. As Kadri Vihvelin eloquently explains:

And we care about our possibilities, both with respect to our immediate choices and with respect to the overall shape of our lives. Looking forward, we want to maximize our potential, develop our abilities as best we can, take advantage of our opportunities, acquire the ability to do the right thing for the right reason, live the best life it is in our power to live. Looking backward, we regret missed or bungled opportunities, blame ourselves for not having tried harder to pursue our dreams, feel guilt for failing to do the right thing we know we should have done, believe that we deserve blame when we fail, without excuse, to live up to our obligations. (2013, 1)

When we do look forward and think about the right plan for our life or look backwards and think about our mistakes, we do not only think about what we actually did (or will do or did), but rather we also think about what we *could have done* or *can* do.

⁷ It is relevant to various discussions about the truth-value of future contingent propositions, but that is of secondary importance for my discussion at this moment. This will come up in Chapter Five though.

Metaphysically, we can capture this intuition with the following principle:

ALTERNATIVE-POSSIBILITIES: A creature *C* is free in ϕ -ing *only if* *C*, when choosing to ϕ , could decide to $\neg \phi$.

This principle is pre-theoretically intuitive and also embedded into our phenomenology of action.⁸ Here, having alternative possibilities is cast as a necessary, but not sufficient condition for having a free will. It is not a sufficient condition because there are possible examples of creatures that have alternative possibilities that are not free. One simple example would be a robot programmed to randomly clap its hands. Or, imagine, a myrmecologist (a scientist studying ants) discover that some aspects of an ant's behavior is random. Neither robots, nor ants, are creatures we paradigmatically think of as free creatures regardless of whether or not they have alternative possibilities.

The second aspect (ii) - that it is our decision (i.e., intention) that determines which outcome comes to obtain- is also important to our intuitive concept of free will. Fischer's illustration of walking down a branching path is a helpful illustration of this point. In the original metaphor, we decide which path to take in the fork in the road. If you simply flipped a coin every time you approached a fork in road, it would not be analogous. In the coin-flipping alternative scenario, even though there are alternative possibilities, you are not making a freely willed decision. There is an intuitive

⁸ To explain this briefly further: just imagine a situation where you only have one option and then imagine a situation where you are choosing *between* options. These are very different kinds of choices and phenomena that do not seem reducible to the other. See also Van Inwagen (2008).

idea associated with free will that when a free person is given an exclusive disjunction of options, *she* is the one who determines which option she takes (although she *could* have chosen either).

In this second aspect (ii), there are really two interesting metaphysical elements. Firstly, there is an assumption that our mental actions really are causal (or make a difference) with respect to what happens. Second, there is an assumption that our decisions and intentions are conscious decisions and intentions. As Gregg Caruso (Caruso, 2012, p. 189) remarks (citing Benjamin Libet and Patrick Haggard (2001)),

A major part of the folk psychology of free will is the belief that *our conscious intentions cause action*. As Patrick Haggard and Benjamin Libet write, “Most of us navigate through our daily lives with the belief that we have conscious free will: that is, we have conscious intentions to perform specific acts, and those intentions can drive our bodily actions, thus producing a desired change in the external world” (2001, 47). This commonsense intuition plays a major role in our sense of free will and is essential to the *up-to-me-ness* that we associate with free will. It is also well supported by phenomenology. In normal cases of voluntary behavior, we experience a conscious intention before the onset of action and naturally take the former to be the cause of the latter. When I switch on my reading lamp, for example, I feel as though it is *I*, my conscious self, that controls the movements of my arms and hands through the conscious formation of goals, intentions, and decisions.

Caruso is commenting on our intuitive or folk notion of free will. He is suggesting, here, that part of the intuitive reason a particular outcome ϕ is up to the agent as opposed to someone or something else is partially explained by the phenomenology of the intention itself: that it is a *conscious* intention.

The way I, and Caruso, are using the term “conscious” follows David Chalmers (1996; 2010). I understand a being to be conscious when there is something that it is like to be that being, and similarly, a mental state (which is a feature of the mind) is conscious when there is something that it is like to be in that state (Chalmers, 2010, p. 104). For example, there is something that it is like to have the perceptual experience of seeing the ceiling of the Sistine Chapel. Sometimes, people use the term “conscious” to denote functional aspects of consciousness and the sense just described. To help avoid this confusion, we can also use the term *phenomenal* consciousness to specifically refer to this aspect of consciousness (Chalmers, 2010, p. 104).⁹ So, when John Wilkes Booth formed an intention to kill Lincoln, there was something it was like to be in that mental state. That is, Booth was consciously aware of this decision. When a person is walking down a branching path and chooses to take the left branch, there is something it is like to make that decision.

Following Terence Horgan (1989), I am going to call this condition Conscious “Quausation” which is a way of saying our actions must be caused by mental states or event in virtue of these states beings mental or conscious states and events. Or,

CONSCIOUS QUAUSATION: A person P has free will *only if* (I) P has the ability to consciously choose an outcome ϕ (i.e., form an intention to ϕ or make a decision to ϕ) where ϕ is distinctly caused by P’s conscious or mental state or event.

⁹ Although, in any mention of “conscious” or “consciousness,” readers can assume I am referring to phenomenal consciousness unless I note otherwise.

According to CONSCIOUS QUAUSATION, there is a privileging of entities or creatures with phenomenal consciousness. For example, a thermostat (which, I assume, does not have phenomenal consciousness) does not and cannot have free will. If at least some animals do not have phenomenal consciousness (as Descartes imagined), then those animals would not and could not have free will (although the question of whether or not animals have phenomenal consciousness is an open question). But, because persons do have phenomenal consciousness, we are the kind of being that would at least *prima facie* qualify for free will.

1.1.2 Formal Definition of Free Will

With CONSCIOUS QUAUSATION and ALTERNATIVE-POSSIBILITIES, I take paradigmatic free will situations to be situations where a person (or other creature with phenomenal consciousness such as a Vulcan) has a decision in front of her and her decision makes a difference (i.e., it distinctly causes) as to what outcome will come about. Following Peter Van Inwagen (1983; 2015), Jessica Wilson and Sara Bernstein (Wilson and Bernstein, 2016; Wilson, 2021), and Christian List (2019), I define free will in the following way:

FREE WILL: A person P has free will with regard to ϕ -ing *if and only if* (I) P has the ability to consciously choose an outcome ϕ (i.e., form an intention to ϕ or make a decision to ϕ)

where ϕ is distinctly caused by P's conscious intention, and (II) when P made a decision to ϕ , P could have chosen $\neg \phi$.¹⁰

In this definition, there are at two jointly sufficient and individually necessary conditions. For what I am going to discuss in this thesis, both conditions will be important, but the alternative possibilities condition (II) will be more of a focus in the subsequent chapters (specifically, Chapter Six). The first condition (I) - CONSCIOUS QUASATION - will be something I focus on in this chapter as it relates to physicalism (§1.3).

It should be clear that I take FREE WILL to be both a metaphysical representation of our intuitive folk belief about free will and a metaphysically interesting doctrine. I should note, though, that any definition of free will is bound to be controversial. However, not defining free will is not an

¹⁰ Compare with the following:

“Let us say that a person's act was free just in the case that (i) that person consciously decided to perform that act rather than some contemplated alternative act(s); and (ii) that person was able to perform at least one of those alternative acts” (Van Inwagen, 2015, p. fn.10).

“Free will, if such there be, involves free choosing: the ability to choose mentally an outcome (an intention to ϕ , or a ϕ -ing), where the outcome is ‘free’ in being, in some substantive sense, up to the agent of the choice” (Bernstein and Wilson, 2016, 1).

“Free will can be understood as a three-art capacity, as has been noted by other philosophers. It consists of [I] the capacity to act intentionally; [II] the capacity to choose between alternative possibilities; [III] and the capacity to [causally] control one's actions” (List, 2019, p. 16).

Bernstein and Wilson both note that different theorists explain what it means for an outcome to be “up to the agent” and one of those options is the alternative possibility explanation. As such, my definition is probably closer to Van Inwagen's definition. We can contrast this with Michael Mckenna and Derk Pereboom's (2016, 6) proposal to understand free will as the “unique ability of persons to exercise the strongest sense of control over their actions necessary for moral responsibility.” I think a more metaphysical definition of free will is preferable to one that essentially deals with moral concepts.

option for the question we are pursuing in this thesis (“Does God have to reach the non-physical shelf to make free creatures?”). However, I want the reader to know that I acknowledge that are objections to FREE WILL. I will engage with some of these objections in Chapter Five. I will discuss CONSCIOUS QUAUSATION further in this chapter and then, again, in Chapter Six (§6.2 and 6.4). For now, I am going to assume that this is a good metaphysical definition of free will that represents our folk notion adequately. Next, I will move to briefly defining determinism, which will be relevant for Chapter Two of my thesis.¹¹

1.2 Determinism

Determinism is a common philosophical thesis.¹² There are different ways to capture the idea of determinism. I am going to explain these ways, and suggest that, in the context of the consequence argument, they are all compatible with my discussion of the consequence argument in Chapter Two.

¹¹ However, because I do not assume physicalism entails determinism, my discussion of determinism is briefer as it is only relevant in Chapter Two.

¹² Following (Hildebrand, 2023, pp. 6-8), it is fairly uncontroversial to state that any plausible candidate for a deterministic law should be at least (1) true, (2) a universal generalization, (3) metaphysically contingent, (4) confirmable by induction, and (5) symmetric (in the relevant sense). Many philosophers believe determinism is true, however there are some detractors. In practice, Hildebrand’s standard is a very high bar for any potential law of nature to meet. This leads some, like Helen Stewart, to suggest that determinism is not plausibly empirically provable:

I therefore do not regard determinism and indeterminism as equal contenders in a battle to which purely empirical science holds the key. I regard them, on the contrary, as deeply *unequal* contenders in a battle to which general metaphysical reflection of a distinctively philosophical sort holds the key- and on the basis of that general metaphysical reflection, I would be completely astounded if physics ever came up with the said scientific evidence (although of course anything is possible). Determinism is, on my view, not a respectable scientific possibility, but a philosophers’ mirage, a position about as likely to be true as the possibility that I am a brain in a vat. (Stewart, 2016, p. 166)

The typical way to capture determinism is to suggest that if one takes any proposition p representing the complete state of the world at one moment t , another proposition q representing the complete state of the world at any other moment than t , which we can refer to as t^* , and a proposition l representing the complete laws of nature of that world, then determinism is the claim that necessarily, p and l together entail q , and necessarily, q and l together entail p (Van Inwagen, 1983, p. 65). So, in a deterministic universe the complete state of the world right now as I eat a burrito (in a conjunction with the laws of nature) entails the complete state of the world at the moment when Julius Caesar crossed the Rubicon. Additionally, the complete state of the world at the moment Julius Caesar crossed the Rubicon (in conjunction with the laws of nature) entails the complete state of the world right now as I eat a burrito.

On this definition, determination is symmetrical according to this definition of determinism; that is, the future entails the past just as much as the past entails the future. The symmetrical aspect of this definition creates one potential objection. The problem is that this conception of determinism likely entails some sort of backwards causation, which is intuitively implausible (although some philosophers are willing to accept the notion).

As a result of this potential worry, some have suggested that we should adopt a narrower definition of determinism. John Carroll and Ned Markosian suggest that we should formulate determinism in a purely future-oriented way; for example, if one takes any proposition p representing the complete state of the world at one moment t and another proposition q representing the complete state of the world at any moment post- t , which we can call t^* (so t is temporally prior to t^*), then determinism can be said to be the claim that p in conjunction with the complete laws of nature entail q (Carroll & Markosian, 2010, p. 50). On this characterization, a proposition expressing the complete state of the world at the moment, say, Julius Caesar crossed the Rubicon in conjunction

with the complete laws of nature would entail the complete state of the world at the moment (right now) where I eat a burrito. However, the backwards entailment would not necessarily hold on this articulation of determinism.

A third potential definition comes from David Lewis (1979) and John Earman (1986). This third conception utilizes modal concepts and is slightly more complicated than the previous two. Lewis claims that:

A deterministic system of laws is one such that, whenever two possible worlds obey the laws perfectly, then either they are exactly alike throughout all of time, or else they are not exactly alike through any stretch of time. They are alike always or never. They do not diverge, matching perfectly in their initial segments but not thereafter; neither do they converge. Let us assume, for the sake of the argument that the laws of nature of our actual world are in this sense deterministic. (Lewis, 1979, pp. 460-461).

In this passage, Lewis is suggesting that the key feature of determinism is that two possible deterministic worlds (w and w^*) with the same laws of nature can never have an identical complete state of the world unless w and w^* always have identical complete states of the world. This is in line with John Earman's thinking of determinism (1986, p. 1). Kadri Vihvelin and Terence Tomkow (Vihvelin & Tomkow, 2016) suggest that we should think of determinism in the following way (building off of Lewis): Determinism is true in any possible world w *iff* the complete set of the laws of nature (l) in conjunction with any proposition p representing the complete state of the w at any time entails a modally unique future and past (with relation to p). In this definition, it is not just that the laws in conjunction with the complete state of the world at one moment entails the state of the world at another moment, it is that the conjunction entails a *modally unique* future

or past. By modally unique, I mean that there are no two possible world (say, w and w^*) such that w and w^* overlap. This would be another way to formulate the deterministic theory.

For my purposes in this thesis, it does not particularly matter which definition is right. There are differences between these formulations of determinism, but they all contain everything needed for the purposes of the consequence argument. In §2.1, I will simply stipulate that the relevant kind of determinism is of the sort Van Inwagen describes (1983) for simplicity. However, I could equally well use either of the other two formulations of the thesis (which is noted in those discussions).¹³

The target of what I will term the P-Modal Consequence Argument, the more traditional version of the consequence argument, is the compatibility of determinism and free will. The target of the B-Modal Consequence Argument – the main argument of this thesis – is the compatibility between a particular kind of *physicalism*, not determinism, and free will. In the next section, I will explain what I mean by physicalism.

1.3 Physicalism

In this section, I will explain what I will take physicalism for this thesis. In §1.3.1, I am going to present physicalism as the combination of two doctrines: Global Supervenience and Nomological

¹³ The only sort of determinism that might not be necessarily compatible with the consequence argument is if someone means *causal determinism* in the sense that every decision has a causally sufficient antecedent. This is what Walter Sinnott-Armstrong takes to be the relevant determinism in the debates over free will and determinism (2022, p. 34). I think this is simply misguided because one is assuming a particular theory of causation without explicitly explaining it. The relevant kind of determinism is a kind of metaphysical determinism, which might entail causal determinism, but causal determinism might not entail metaphysical determinism.

Closure. In §1.3.2, I will explain that, for my purposes, there are three relevant versions of physicalism: Type A (eliminative or reductive) physicalism, Type B (identity) physicalism, and Type B (non-reductive) physicalism. I take it that Type-A physicalism and Type B (identity) physicalism both struggle to explain free will and Type B (non-reductive) physicalism is plausible sufficient or compatible with free will.

1.3.1 Global Supervenience, Fundamental Physicality, and Nomological Closure

Barry Loewer characterizes physicalism as the claim “that all facts obtain in virtue of the distribution of the fundamental entities and properties – whatever they turn out to be – of completed fundamental physics” (Loewer, 2001, p. 37). So, roughly, physics (of the future and complete kind) should explain everything. However, there are several nonobvious aspects to Loewer’s characterization. I am going to begin by discussing the claim of *fundamentality*.

On the physicalist view of the world, we can categorize the world into (I) facts concerning fundamental entities and properties and (II) facts concerning non-fundamental entities and properties. Typically, we think of non-fundamental entities and properties as macrophysical entities and properties. By macrophysical, I mean the phenomena we experience in everyday life, such as an entity like the Statue of David and its property being worth 82 million dollars. We can characterize macrophysicality as the broad level of description above what we may refer to as the fundamental physical level of reality. The fundamental physical level is usually called the microphysical level. By fundamental microphysicality, I mean the phenomena at the level of reality for which (I) there is no more reduction (i.e., there are no mereological simpler entities and features) and (II) on which everything else metaphysically or logically depends according to the

physicalists. Our current microphysics describes objects like quantum fields or elementary particles and these objects have features such as a field's amplitude or an electron's spin.¹⁴

One might get the impression that physicalism is an empirical or *a posteriori* theory (e.g., “everything either is or depends on *whatever* the future complete physics says there is fundamentally”). However, this would be a misunderstanding of how most physicalists understand physicalism.¹⁵ The physicalist understanding of fundamentality has an *a priori* limitation on what the future complete physics can look like in order for physicalism to be true. Physicalists define what fundamental physical properties are by pointing to the kinds of properties necessary for explaining “non-sentient objects” like chairs or mountains (Jackson, 1998, p. 8). So, the *a priori* constraint is that in order for the complete future physics to be compatible with physicalism, it must not postulate distinct consciousness or conscious-like objects, properties, or explanations for fundamental microphysicality. For example, a physicalistically acceptable complete physics can postulate a new kind of quantum field or some kind of localizing property (i.e., a local beeable) as being part of fundamental physics, but it cannot postulate that proto-consciousness is a fundamental property of its own.

¹⁴ It is true our physics is not yet complete, so it is theoretically possible that there is something further down than the particles and fields our current physics postulates.

¹⁵ There are some that would understand physicalism in this way (Dowell, 2006).

Without this kind of *a priori* constraint physicalism, *a posteriori* physicalism would be conceptually consistent with dualism and other (currently) non-physicalist theories. Philip Goff (2017, p. 27) expresses this worry in the following way:

Suppose future physicists postulate fundamental mentality, psychic powers, or even souls. It would then turn out that the truth of physicalism was consistent with the existence of fundamental mentality, which many philosophers find deeply counterintuitive. Physicalism is supposed to be the contemporary analog of seventeenth-century materialism, the arch opponent of dualism and idealism. If physicalism turns out to be consistent with such views, then it seems to lose its point.

This is why the physicalism should be thought of as having an *a priori* constraint on what can constitute fundamental physicality in the definition of physicalism. The constraint would simply be that this futuristic complete physics *cannot* postulate any distinct conscious entities, properties, or explanations as part of fundamental microphysicality (in order for physicalism to remain true).

However, it is important to note that physicalism does not necessarily deny the existence of consciousness or minds (or other macrophysical properties and objects). Presumably, there would be mental entities or properties listed as non-fundamental physical entities and properties in this physicalistic, future complete physics book (just like hurricanes or chairs would be non-

fundamental entities), or consciousness would be *identified*, in some manner, with some fundamental physical phenomena.¹⁶

In Loewer's characterization mentioned at the start of this discussion he said "all facts obtain in virtue of" these fundamental facts of the kind just discussed (Loewer, 2001, p. 37). In this definition there is a key explanatory relation ("in virtue of") in his characterization of physicalism. A more formal, and traditional, definition of physicalism can help us understand this "in virtue of" explanatory relation, and this formal definition comes from Frank Jackson (1998, 12).¹⁷ According to him, "[p]hysicalism is [the claim] that if you duplicate our world in all [fundamental] physical respects and stop right there, you duplicate it in all respects" (ibid.). Alternatively, using the possible world language:

GLOBAL SUPERVENIENCE: Physicalism is true in our world *only if* "[a]ny world which is a minimal [fundamental] physical duplicate of our world is a duplicate simpliciter of our world" (Jackson, 1998, p. 12).

It is important to realize what GLOBAL SUPERVENIENCE is not suggesting: GLOBAL SUPERVENIENCE is not suggesting there are no possible worlds with non-physical entities (e.g., worlds with ghosts). Additionally, it is also not suggesting that there are no possible worlds with identical physical facts as our own but with non-physical entities also added. For example, if physicalism is true in the

¹⁶ For example, perhaps some sort of realizer-functionalism (Lewis, 1966) ends up being true. If this is the case, then a conscious state is identified with the token fundamental physical realizer state, but there is not a distinct conscious state in fundamental physicality.

¹⁷ Loewer (2001, p. fn. 4).

actual world w , it is still possible for there to be a possible world w^1 that is a duplicate of w except for that there is additionally an epiphenomenal ghost present in w^1 .

Rather, GLOBAL SUPERVENIENCE is suggesting that for physicalism to be true in our world, the fundamental microphysical facts (i.e., truths concerning the fundamental microphysical objects, properties, and laws) of our world logically or metaphysically entail *all* the positive facts of our world. David Lewis (Lewis, 1986, p. 14) gives an example of this dependence relationship with a dot painting of a bird, which is a helpful analogy.¹⁸ These sorts of paintings have various “macro” level properties (e.g., looking like a bird) and “micro” level properties (e.g., the dots in a certain pattern). The paintings depend on the dots in the same manner as macrophysicality is supposed to depend on the (fundamental) microphysicality. Of these pictures, Lewis points out, “no two pictures could differ in their [macro] level properties without being different, somewhere, in whether there is or there isn’t a dot” (ibid.). Daniel Stoljar explains how this helps our idea of physicalism, “[t]he basic idea is that the [fundamental] physical features of the world are like the dots in the picture, and the psychological or biological or social features of the world are like the [macro] properties of the picture. Just as the [macro] features of the picture [depend] on the dots, so too everything [depends] on the physical, if Physicalism is true” (Stoljar, 2022, p. S2.1). This dot analogy is helpful because if one imagines *duplicating* the dots themselves on a different canvas, one will also imagine the image the dots create. It is impossible to have the dots in their exact state when they undergird the bird painting and *not the bird!* Regardless of whether the

¹⁸ Although, his example is just a dot matrix, I do not think he used a bird specifically.

dependence relation is logical or metaphysical, *all* positive facts about the actual world must *depend* on the fundamental microphysical facts of the world if physicalism is true in the same way a bird painting depends on the dots. This dependence relation is the second critical aspect of Loewer's characterization.

One might simply define physicalism in terms of GLOBAL SUPERVENIENCE. However, physicalists do all agree that one additional doctrine is worth including in the definition of physicalism, a doctrine known as NOMOLOGICAL CLOSURE (this doctrine is also often referred to as *causal* closure.)

Multiple authors have articulated a closure-type doctrine in various ways, for example¹⁹:

- (I) If a physical event has a cause that occurs at *t*, it has a physical cause at *t* (Kim, 2005, p. 15)
- (II) All physical effects have sufficient physical causes (Papineau, 1998, p. 375)

These formulations express a similar idea. Both are all trying to suggest that non-physical entities cannot cause physical effects, but (I) and (II) are not precisely equivalent. And, following Lowe (2000, pp. 574-576) and Gibb (2015, pp. 628-629), I would suggest that these articulations are not ideal articulations of the core physicalist idea. Kim's version does not specify in great detail what metaphysical theory of causation is sufficient or necessary for Closure. For example, if you hold

¹⁹ For a more extensive survey of prospective causal closure formulations see Lowe (2000) and Gibb (2015). Discussing every variation and articulation of this doctrine is not very beneficial for my purposes here.

to a counterfactual view of causation, it is difficult to maintain closure on this articulation.²⁰ Papineau's articulation presupposes the notion of causal sufficiency and, thus, some sort of causal determinism. It is quite possible that contemporary quantum theory is not compatible with such a definition. For example, on many interpretations, the location of an electron is not sufficiently caused by its prior position and dynamic properties. This seems problematic without resorting to any metaphysical arguments. In general, it seems appropriately prudent to articulate the closure doctrine in a way that would be compatible with both indeterministic and deterministic models of causation.

Following Loewer (2015, p. 41), I therefore define the closure doctrine in the following formulation:

NOMOLOGICAL CLOSURE: for any fundamental microphysical event (except the ones at the initial conditions), the objective probability of that event obtaining is identical to their probability solely given prior microphysical events of the universe in conjunction with the fundamental laws of nature.

On this version of the doctrine, the possibility that the world is indeterministic is left open. The probability of an event could be .6 or 1, but this would not matter. What NOMOLOGICAL CLOSURE is asserting is that the probability of E (where E is, say, an electron spinning upwardly in a given direction) given all antecedent factors (i.e., the objective probability) is x , and x must be identical

²⁰ I will point this out with more detail in §4.2-4.3, as this is a common rebuttal to arguments using this version of Closure. Additionally, see Loewer (2015).

to the probability to E only given the antecedent microphysical factors in conjunction with the laws of nature. NOMOLOGICAL CLOSURE also leaves open which theory of causation is correct.

One possible misconception about NOMOLOGICAL CLOSURE is that it means that the non-fundamental physical objects and properties would be epiphenomenal. That is not the case. One benefit of NOMOLOGICAL CLOSURE is that it does not necessarily invoke causal notions at all. It only involves nomological aspects of *fundamental* microphysical objects and properties (e.g., the state of a group of electrons). That is a more limited doctrine than many other options.

To review, in this thesis I will understand physicalism as the conjunction of GLOBAL SUPERVENIENCE and NOMOLOGICAL CLOSURE. I assume that any genuine version of physicalism must be committed to these two doctrines (or some equivalent form of them). However, physicalists do disagree internally about various questions. Next, I consider two doctrines that physicalists disagree on: Reductionism and Mental Ontological and Causal Irreducibility.

1.3.2 Mental Ontological and Causal Irreducibility

Physicalism is supposed to be a theory about reality, but also, specifically, about consciousness (i.e., phenomenal properties or states). Following David Chalmers, I understand a phenomenal property or state to be a first-person perspective property like ‘having the perceptual experience of

seeing the Sistine Chapel.’ If GLOBAL SUPERVENIENCE is true, phenomenal properties depend on fundamental microphysical properties like other non-fundamental physical properties.²¹

As it pertains to phenomenal consciousness and physicalism, one famous problem is Chalmers’ Hard Problem of Consciousness (1996; 2010). This problem challenges one to “explain how and why physical processes give rise to phenomenal consciousness” (Chalmers, 2010, p. 105). Roughly speaking, we can offer physical explanations for different kinds of behaviors and functions fairly well, but it is hard to see how properties like mass, time, and space can give us qualitative properties like phenomenal consciousness, creating an explanatory gap. This explanatory gap is supposed to create a metaphysical and modal difference between phenomenal consciousness and fundamental physicality.

Chalmers puts this argument formally in the following way: Chalmers uses “ P [as] the conjunction of all microphysical truths about the universe, and Q , an arbitrary phenomenal truth about the universe. (Here, ‘&’ represents ‘and’ and ‘ $\sim$ ’ represents ‘not.’)” (2010, pg., 108). Using these symbols Chalmers’ (2010, pg., 109) puts his Hard Problem in the following way:

1. It is conceivable that $P \& \sim Q$.
2. If it is conceivable that $P \& \sim Q$, it is metaphysically possible that $P \& \sim Q$.

²¹ Phenomenal properties presumably also depend on neural properties. However, physicalism is not the thesis that mental properties depend on neural properties, but rather that all properties (e.g., phenomenal properties) depend on fundamental microphysical ones. Therefore, phenomenal properties must depend on fundamental microphysical properties as well as the neural properties they depend on.

3. If it is metaphysically possible that $P \& \sim Q$, then materialism is false.

4. Materialism is false.

In this argument, one way to conceive of $P \& \sim Q$ is to imagine of an entity that Chalmers' calls a zombie. A (philosophical) zombie is a physical entity that is at least microphysically identical to an ordinary human being, except that a zombie lacks phenomenal consciousness. An entity that has all the same physical bits as myself (e.g., the same configuration of electrons, atom, and molecules), but is completely dark inside its subjective self. My personal zombie would be like me physically, but have no first-person experience (i.e., phenomenal consciousness).²² There would be nothing it is to be like a zombie.

To put this concretely, imagine I am enjoying gazing at the roof of the Sistine Chapel, having a particular perceptual experience. If my zombie twin were next to me, he would not be in this state. He would be physically just like me, even looking at the Chapel and possibly reporting the same things as me, but there is nothing it is like to be a zombie. Although, the zombie would function and behave identical to myself (it might report the same "wow" as I do for example).²³

²² Another way to run the argument would be by conceiving of a hypothetical being called an invert, which is microphysically, functionally, and behaviorally identical to an ordinary human being, but the invert has different phenomenal experiences than the original ordinary human being (Chalmers, 2010, pg., 108).

²³ My invert twin would be having a phenomenal experience, but perhaps he is in a different state than I am in (because blue is perceived as red to him or some other slight difference) even though we have identical physical states.

When we imagine or conceive of a zombie, it is important that this claim is precise because there are different sense of conceivability. Chalmers notes six different possible aspects of conceivability (prima facie, ideal, primary, secondary, positive, and negative). According to Chalmers, “we can say that a sentence *S* is conceivable when *S* expresses a coherent hypothesis: one that cannot be ruled out a priori” (Chalmers, 2010, p. 143). For example, “rockets can fire rocket launchers” is conceivable in this sense. Additionally we can distinguish two kinds of further claims: prima facie and ideal conceivability. Firstly, “[w]e can say that *S* is *prima facie* conceivable (for a subject) when that subject is unable to rule out the hypothesis expressed by *S* by a priori reasoning on initial consideration” (emphasis added, *ibid.*). Secondly, “we can say that *S* is *ideally* conceivable when the hypothesis expressed by *S* cannot be ruled out a priori even on ideal rational reflection” (*ibid.*). For example, $x^{bc} = (x^b)^c$ is prime facie conceivable, but is not ideally conceivable (because it is mathematically false). Golbach’s conjecture would be another good example of something that is certainly prima facie conceivable, but it controversial as to whether it is ideally conceivable. Thus, one distinction we can make is between something being *prima facie* conceivable and ideally conceivable.

Another kind of important distinction in varieties of conceivability is between positive and negative conceivability. *S* is positively conceivable if *S* is imaginable. For example, Chalmers’ zombies are positively conceivable because you can picture them in your mind. There is a thing there that you can actually imagine. Contrasted with this, negative conceivability is the merely the idea that one entertain *S* because one is “unable to rule it out” (*ibid.*, 144).

There is one final notable distinction in conceivability claims. Sometimes we talk about *S* being “prima facie conceivable to one without relevant empirical knowledge” (*ibid.*, 145). In the famous Kripkean example, the claim that “water is not H_2O ” seems conceivable, but once one has the

relevant knowledge of actual chemistry, it is not conceivable. If we use empirical reasoning to help our conceivability claim S, then we can say that S is secondarily conceivable. In this sense, “water is not H_2O ” is not ideally secondarily conceivable because water in our world is H_2O . However, primary conceivability is the kind of conceivability we use in exclusively a priori reasoning. Chalmers defines primary conceivability in the following way: “We can say that S is *primarily conceivable* (or *epistemically conceivable*) when it is conceivable that S is *actually* the case. We can say that S is *secondarily conceivable* (or *subjectively conceivable*) when S conceivably *might have been* the case” (Chalmers, 2002). He goes on to explain (ibid. pp. 151-152):

Primary conceivability is grounded in the idea that for all we know a priori, there are many ways the world might be. The oceans might contain H_2O or they might contain XYZ; the evening star and the morning star might be the same or distinct; and so on. We can think of these ways the world might be as *epistemic possibilities*, in a broad sense according to which it is epistemically possible that S if the hypothesis that S is not ruled out a priori. When S is epistemically possible, there are usually a number of imaginable situations such that if they actually obtain S will be the case. These situations can be taken to verify S, when they are considered as actual... Secondary conceivability works quite differently. It is grounded in the idea that we can conceive of many counterfactual ways that the world might have been but is not. When we consider imagined situations as counterfactual, we consider and evaluate them in the way that we consider and evaluate counterfactual possibilities in the subjunctive mode. That is, we acknowledge that the character of the actual world is fixed, and say to ourselves: if the situation *had* obtained, what *would have been* the case? If we judge that had the situation obtained, S would have been the case, then we judge that the situation verifies S when considered as counterfactual.

The claim that zombies are conceivable in the Hard Problem is really that they are *ideally positively primarily conceivable*. That is, Zombies are entities we can (i) actually imagine, (ii) are imaginable in a purely *a priori* manner, and (iii) and are imaginable upon reasonable rational reflection.

From the conceivability of $P \& \sim Q$ (or zombies or inverts), we are supposed to be able to infer that $P \& \sim Q$ (or zombies or inverts) is *metaphysically possible*. As Chalmers' (2010, pg., 108) notes,

Zombies are probably not naturally possible: they probably cannot exist in our world with its laws of nature. But the argument holds that zombies *could have* existed, perhaps in a very different sort of universe. For example, it is sometimes suggested that God could have created a zombie world if he had so chosen. From here, it is inferred that consciousness must be nonphysical. If there is a metaphysically possible universe that is physically identical to ours but that lacks consciousness, then consciousness must be a further, nonphysical component of our universe. If God could have created a zombie world, then (as Kripke puts it) after creating the physical processes in our world, he had to do more work to ensure that it contained consciousness.

There is a modal difference between P and Q , Q can possibly obtain without P . If God had to create *more* than fundamental microphysicality in order to create consciousness (as we know consciousness in the actual world), then GLOBAL SUPERVENIENCE (and thus physicalism) turns out to be false.

This is a controversial argument, but it does help us differentiate a couple of physicalist views by their responses to it. There are two types of responses that I will consider first: Type-A physicalism (i.e., the eliminativist strategy) and Type-B (identity) physicalism. Both of these strategies, I argue,

either struggle to overcome Chalmers' Hard Problem, or struggle to potentially explain free will sufficiently.

The first kind of response comes from what Chalmers calls Type A materialism (or physicalism). The Type-A physicalist denies that zombies or invert are conceivable (in the relevant conceivability sense mentioned above). Because they are not conceivable, they [zombies] are not metaphysically possible, and, hence, a threat to physicalism. This would be the "view that on reflection there is nothing in the vicinity of [phenomenal] consciousness that needs explaining over and above explaining the various functions: to explain these things is to explain everything that needs to be explained" (Chalmers, 2010, pp. 112-113). In an example where my zombie twin and I are staring the ceiling of the Sistine Chapel, both I and my zombie can access memories and report information (as example functions). The Type-A physicalist would be suggesting that explaining these sorts of functions or behaviors would be the only thing the physicalist needs to do.²⁴ The phenomenal data that we believe we experience is not actually what it appears to be.

This view is counterintuitive in some clear respects. As Chalmers notes, "it appears to deny the manifest" (2010, pg., 113). There are clearly functional phenomena that require explanation: I can access certain memories and report those to someone who asks, for example. Accessing information and reporting that information are kinds of functional or behavioral states. However, I also have a further phenomenon that, as it seems to me, requires explanation: phenomenal

²⁴ The Type-A physicalism can take the form of eliminativism in the sense that the claim would be that phenomenal properties are not genuine properties full stop. Although, on the Lewisian view, they would be properties like 'being grue' is a property. It can also take the form of identifying phenomenal states in terms of functional or behavioral states, which is a sort of reductionism.

phenomena. As Chalmers' notes, to "flatly deny the further [phenomenal] truth, or to deny without argument that there is a hard problem of consciousness over and above the easy problems, would be to make a highly counterintuitive claim that begs the important questions" (ibid). Now, as Chalmers notes, sometimes counterintuitive claims end up being true, but it seems like the critic should be able to produce an argument for why zombies are not really conceivable (in the relevant sense).

For my purposes, this sort of response is even more problematic. This thesis is interested in the potential compatibility/incompatibility between free will and physicalism. In order for physicalism to be compatible with free will, it must be possibly able to satisfy the CONSCIOUS QUAUSATION conditions. But, by denying that phenomenal consciousness is a genuine or distinct phenomenon to be explained, it would inevitably deny CONSCIOUS QUAUSATION. In other words, our experience of our free will is that we, via our conscious intentions, cause our physical movements and various outcomes. If one can explain only the functional or behavioral aspects of consciousness without any distinct reference to phenomenal aspects, then CONSCIOUS QUAUSATION is false.

Other physicalists respond to the zombie argument in a different manner. Chalmers calls this second group Type-B physicalists. Chalmers differentiates two approaches within this category, but calls both approaches those of Type-B theorists. For clarity I will refer to one group as Type-B *identity* physicalists and another group as Type-B *non-reductive* physicalists. I differentiate between these groups because they make relevantly different metaphysical claims about phenomenal properties. Both groups would argue that zombies are conceivable (in the relevant sense), but rejects the notion that zombies are metaphysically possible.

Type-B (identity) physicalism holds that phenomenal properties are identical to fundamental microphysical properties. By virtue of this, this view would hold that all phenomenal properties

are genuine properties, they just reduce down to microphysical properties. However, on this view, there is a conceptual or epistemic gap between physical or phenomenal truths (although there is not metaphysical or ontological gap because both properties are identical to one another).

Often proponents of this view argue the relation between the phenomenal and physical would be similar to the *a posteriori* identities between Clark Kent and Superman, H₂O and water, or genes and DNA (Block & Stalnaker, 1999; Seale, 1992; Loar, 2017). As Chalmers (Chalmers, 2010, p. 115) explains,

The concept *water* is different from the concept H₂O, but they are found to refer to the same thing in nature. On the type-B view, something similar applies to consciousness: the concept of consciousness is distinct from any physical or functional concepts, but we may discover empirically that these refer to the same thing in nature. In this way, we can explain why there is an epistemic gap between the physical and phenomenal domains while still denying any ontological gap. This yields the attractive possibility that we can acknowledge the deep epistemic problems of consciousness while retaining a materialist worldview.

The idea here is that Louis Lane may know a lot about Clark Kent, but may not be able to discern *a priori* everything about superman because she does not know (empirically) that Clark Kent is superman. Similarly, *I* should know that it is impossible to have water in my glass, but not H₂O in my glass, *but* a 13th century Viking may not be able to know this because he does not know the requisite chemistry facts. In this manner the concept of water or superman is not the same epistemic concept as H₂O or Clark Kent, but these concepts do actually refer to the same properties. These sorts of properties are additionally interesting because there are related

In the sciences, there are well-known gap problems between properties like being water and being H₂O, that is, properties at different *levels* of scientific description (or different sorts of scientific laws like physics, chemistry, biology, and so forth). Take the temperature in the room I am in right now (where the temperature is at 70 degrees Fahrenheit). The property of being 70 degrees (D) is no doubt instantiated co- temporally and co-materially with an arrangement of particles in a room. Let us call this particular arrangement Σ . The room having the property of Σ and D are certainly closely related, in fact some may not be any ontologically difference between them at all. However, D certainly appears to have certain properties that Σ does not. For example, D (the property of being 70 degrees) could have been instantiated with an arrangement of particles like Σ , but where, for example, the placement of a potassium ion was different. Or, we could have had different particles entirely. Regardless of how we want to explain these apparent differences, empirically, many Type-B theorists would argue that our best empirical theories posit identities between these sorts of phenomena, even with an apparent gap at times.²⁵ So, Type-B (identity) physicalists would argue that there are these *a posteriori* identity relations that we already accept, and, by analogy, phenomenal consciousness is in a similar *a posteriori* identity relation with some microphysical state (Block & Stalnaker, 1999; Seale, 1992).²⁶ Using this strategy, one would argue that zombies are conceivable, but not metaphysically possible.

²⁵ I think that the multiple realization argument is a good reason in of itself to reject reductionism of this kind. However, I am merely presenting what this sort of view argues here.

²⁶ Chalmers (2010, pp. 305-336) does a good job overviewing various Type-B views. One prominent alternative family of Type-B views comes from Brian Loar (2017), and is often referred to as the phenomenal concept strategy.

A second-related- strategy is inspired from Brian Loar famous article “Phenomenal States” (1990/2017), which argues that there is something unique about the gap between the phenomenal and the physical, which is explained by the unique nature of the phenomenal concept itself. Loar argues that phenomenal concepts are *recognition concepts* where recognition concepts are “concepts deployed when we recognize an object as being one of *those*, without relying on theoretical knowledge or other background knowledge” (Chalmers, 2010, p. 309). Chalmers uses the example of recognizing a sort of cactus as one that causes a particular sort of experience, such as the experience of pain when you sit on it (ibid). Phenomenal concepts are ones in which we use to refer to those experiences directly. There are other versions inspired or related to Loar’s proposal, but the essence is that phenomenal concepts can be explained a unique manner that self-explains why zombies are conceivable, but not metaphysically possible. If one uses some version of this strategy, then zombies are not metaphysically possible by definition, but are conceivable.

Following Lewis (1984) and, especially, Loewer (2015), we can distinguish properties in two ways: Firstly, properties can be fundamental or non-fundamental (As discussed previously). However, secondly, properties can also be either genuine (i.e., G-properties) or non-genuine (i.e., non G-properties), but not both. This second category is important to why I think Type B (identity) physicalism is incompatible with CONSCIOUS QUAUSATION.

According to this view, there is something unique about the gap between phenomenal consciousness and neural states, but this can be explained by the nature of phenomenal consciousness.

On Lewis' theory of properties, predicates are typically properties (1984; 2002; 2004).²⁷ (Lewis also thinks of properties as classes or sets, but Lewis' particular metaphysics of properties is not necessary to my general conception about genuineness.²⁸) And, on Lewis' view, some properties are explanatorily superior to others; by this, I mean they fit into causal or nomological explanations. In Lewis' nomenclature, we can call this explanatory superiority, *naturalness*. A property can be more or less *natural* than another if it is more or less explanatorily useful. And, for Lewis, the *most* natural properties are what he calls the "perfectly natural" properties, which correlate to what I call fundamental properties (e.g., the distribution of electrons or the amplitude of a wave-function).

An example of a non-fundamental, but fairly natural, property would be an emerald's property of being green. The greenness of an emerald is relevant for various kinds of causal and nomological explanations. For example, Sally purchases an emerald rather than a diamond *because it is green*. The color would be relevant for lawful explanations of light reflecting or going through the gem. Additionally, learning that emeralds are green can help us categorize them with alike objects and exclude dissimilar objects for various purposes.

²⁷ Weatherson (2021) shows that in Lewis' later papers, he "notes that some predicates don't correspond to properties or relations. There is no property of being non-self-instantiating, for instance, though we can predicate that of many things."

²⁸ Weatherson (2021) explains that on Lewis view properties are determinative of sets of possible individuals and the contents of our beliefs. Thus, to believe something is to "locate yourself within a class of possible individuals; to believe that you are one of the individuals with the property."

There are other kinds of predicates that may be properties, but are not as natural as greenness. Emeralds have the interesting fact about them that, as they develop, they are green for a while and then become blue. So, emeralds are “either green and observed before a certain time t or blue and not observed before t ,” famously also called “being grue” (Goldman, 1954). ‘Being grue’ is an odd disjunctive predicate, but it is understandable. Because it is a predicate, on the Lewisian view, we can understand being grue as a property: grueness. Grueness, though, is less natural, less explanatorily useful than greenness.

So, on this general Lewisian conception, grueness and greenness are both non-fundamental properties (whereas being an electron is a good candidate for a fundamental property). However, we need to make a further distinction, following Loewer (Loewer, 2015, p. 44), we can distinguish a subset of properties as *genuine* properties or G-properties in the following way:

The distinction between properties that are either fundamental or law involving and those that are not is the distinction that most philosophers have in mind when they speak of some properties as “real” and others as not. Some plausible candidates for G-properties are *positive charge*, *being a gas*, *mutation rate*, *episodic memory*, and *being a monetary exchange*. Plausible candidates for not being G-properties are *being grue*, *being postmodern*, and *being a gas or a mutation*

By “law involving” (ibid), Loewer means any property relevant for “a simple true generalization or equation that is counterfactual supporting, projectible, sufficiently simple, and so on” (ibid.). So, being grue would not be a great candidate for a physicalistically acceptable category of G-properties because it is so unnatural, being a gas or a mutation, likewise, would not fit into any lawful or causal explanation well.

Here is some terminology (again, following Loewer (2015, pp. 44-45): Let us call all *genuine* properties, which are properties that are instantiated in the real world and are lawful or causal, G-properties (i.e., distinct genuine properties) and let us call all fundamental properties F-properties. Physicalists of all stripes would contend that all F-properties in the actual world are G-properties.²⁹ Additionally, physicalists of all stripes would contend that the properties picked out by the natural sciences (e.g., biology, chemistry, physics, etc.) are G-properties as well.³⁰ Type-B (identity) physicalists and Type-B (non-reductive physicalists) can be distinguished from each other in that, for the identity theorist, if any phenomenal (or mental) property *is a G-property*, then it must be a F-property or, at least, a property picked out by a natural science like physics or biology (Loewer, 2015, p. 45). (The same reasoning would apply to other special science properties like being a monetary exchange or being a hurricane.) That is, the mental property must be identical to a neurophysiological, biological, or fundamental physical property.³¹

As I mentioned before, the Type-B (identity) physicalism has a unique way of overcoming the Zombie Argument, by accepting the conceivability of zombies, but denying the metaphysical possibility of zombies. Unlike Type-A views, Type B views (of this family) do want to offer an explanation of phenomenal consciousness. But, these sorts of explanations will not produce a sort

²⁹ Loewer uses the term “P-property” for fundamental physical property. I think that is confusing so I use “F-property” (ibid).

³⁰ These would presumably be reducible or identified with F-properties.

³¹ Of course, if it was identical to a neurophysiological property, that neurophysiological property would be reducible itself. A prominent view in this neighborhood would be a view called *realizer functionalism*. According to this view, the higher-level special science concept merely provides a description of the physical or fundamental functional state. So, a pain-state, for example, helps identify some microphysical state *as that pain-state*.

of phenomenal state that is *distinctly causal*. To be distinctly causal is to be a G-property, but phenomenal concepts or properties are not themselves distinct G-properties.³² These views will say mental properties are causal, or are G-properties, because they are neurophysiological or F-properties (just for whatever reason, there is some conceptual or epistemic gap like Louis Lane with Clark Kent/Superman). A problem for these identity views is that it is hard to see how this view could be compatible with CONSCIOUS QUAUSATION anymore than Type-A views could be.

Type-B views, though, do not have to take this identity position. As Loewer says, “[The Type-B identity theorist] says that all of nature’s joints are physical, whereas [The Type-B non-reductive physicalist] says that some are mental” (ibid). In order for there to be genuine *non-fundamental* (mental) properties in a physicalist world, there must a genuine non-identity relation between fundamental microphysical properties and phenomenal properties (and possible other properties). To explain this view, I need to explain two different non-identity views: constitution and grounding relations.

The first relation is the constitution relation, a mereological view concerning the part-whole relationship. For example, consider the relation between the lump of clay (part) to the Statue of David (whole). Constitution theorists argue that the Statue of David and the lump of clay exist

³² Here, I acknowledge that some Type-B theorists might think there is a way to have this kind of physicalism and free will. It is an interesting question by itself. However, I think the identity relation makes free will very difficult to make coherent under the view of free will I posited earlier. Additionally, I think most physicalists would prefer a non-identity relation. For example, even Alyssa Ney, who sometimes refers to herself as a reductionist, advocates for a non-identity relation (Ney, 2021).

simultaneously in the same space-time region.³³ Therefore, these two entities share the same matter and parts. However, the Statue of David contains several properties – temporal, modal, and aesthetic – not shared by the lump of clay. For example, let us assume Michelangelo sculpts the Statue of David on Tuesday. If that is true, then the Statue of David could not have existed on Monday; however, the lump of clay did exist on Monday. That is an example of a non-identical property between the two objects (i.e., the statue could not have existed on Monday but the lump did (and therefore could have) exist on Monday). Alternatively, consider the following modal properties: It is true that if the Statue of David had been sculpted from a different token lump of clay or a different type of building material (e.g., a lump of marble), it would have still been the Statue of David. However, the lump of clay cannot be a different type of material (and this particular lump cannot be another particular lump). Therefore, this is another non-identical property the Statue has, which the lump does not. Therefore, assuming a basic law of identity (where two entities are identical *iff* they have the exact same properties), the lump and Statue are non-identical objects. However, it is evident that the Statue of David is also ‘nothing over and above’ the lump of clay; it is not as if Michelangelo needs the lump of clay plus some other ingredient to make the Statue of David. In this manner, some physicalist theorists argue that configurations of fundamental microphysical objects and properties *constitute* non-fundamental objects and features (e.g., the Statue of David).

³³ Lynn Rudder Baker (2009) and David Wiggins (1968; 1980) are two such theorists.

Another possible non-identity relation is the grounding relationship. Grounding relationships are primarily explanatory relationships. For example, the proposition “a party is occurring” is true in virtue of, but is not identical to, the proposition that “Jussi, Yujin, and Ben are celebrating.”³⁴ These two propositions are not identical propositions because if Yujin left the celebration and Henry replaced him in the celebration, then the proposition “a party is occurring” would still be true.

The potential worry for either the constitution or grounding relationship (in the context of combining one with physicalism) is that the existence of non-fundamental objects and properties might morph physicalism into emergentism.³⁵ Regardless of whether one prefers constitution or grounding (or some other non-identity relation), Umut Baysan (2019, 79) puts forward the following plausible conditions on any physicalist non-identity theory such that this entailment relation will be compatible with GLOBAL SUPERVENIENCE and NOMOLOGICAL CLOSURE (which he calls “realization”):

1. Fundamentality or Priority: P realizes M iff P is more fundamental than M (i.e., biological properties are more fundamental than botany properties; properties in physics are more fundamental than meteorological properties).
2. Metaphysical Entailment: P realizes M iff P metaphysically necessitates M

³⁴ I believe Philip Goff uses an example like this in his (2017, C. 2), which inspired this example.

³⁵ The worry is that if we suggest that tables and chairs are genuine non-fundamental properties and objects that have distinct causal and nomological powers, then nomological closure might be violated, thus violating physicalism. Umut Baysan discusses the difference between emergentism and non-reductive physicalism well in his article (Baysan, 2019).

3. Causal Derivation: P realizes M iff all of M's causal powers derive from P's causal powers.
4. Non-Identity: P realizes M iff P is not identical to M in either a token or type-identity sense.

On Baysan's view, as long as the non-identity relation (4) *also* has the (1)-(3) components, it should be sufficiently physicalistically acceptable.

So, physicalists can accept GLOBAL SUPERVENIENCE and NOMOLOGICAL CLOSURE consistently with the following:

ONTOLOGICAL AND CAUSAL IRREDUCIBILITY: Some non-fundamental objects and properties are (I) *not identical* to the configurations of fundamental microphysical objects and properties they metaphysically depend on and are (II) distinctively causally efficacious with respect to the configurations of fundamental microphysical objects and features they depend on (see Wilson, 2021, pp. 1-2).

Here, I do intend "properties" to be read broadly to include processes, relations, states, and events. So, ONTOLOGICAL AND CAUSAL IRREDUCIBILITY is expressing the view that objects like hearts are not necessarily identical to a configuration type or a token configuration of microphysical objects and properties. A heart is realized by a token realizer configuration, and has distinct causal powers. Hearts can feature in anatomical and biological causal explanations in a way that purely microphysical explanations cannot.

But, of course, physicalism in mind is supposed to be a thesis, not only about the metaphysical nature of reality, but about phenomenal consciousness more specifically (as discussed in §1.3.2). If GLOBAL SUPERVENIENCE is true, then phenomenal properties depend on fundamental microphysical properties like other non-fundamental physical properties too. Phenomenal properties presumably also depend on neural properties. However, physicalism is not the thesis

that mental properties depend on neural properties, but rather that all properties (e.g., phenomenal properties) depend on fundamental microphysical ones. Therefore, phenomenal properties must depend on fundamental microphysical properties as well as the neural properties they depend on.

Physicalists who accept ONTOLOGICAL AND CAUSAL IRREDUCIBILITY have a different way of handling the questions about phenomenal consciousness than Type-A (reductive), Type-A (eliminative), and Type-B (identity) physicalists. As David Chalmers rightly notes:

There is another line [the Type-B physicalist] can take. One can first note that an identity between consciousness and physical states is not strictly required for a [physicalist] position. Rather, one can plausibly hold that [physicalism] about consciousness simply requires that physical states necessitate phenomenal states in that it is metaphysically impossible for the physical states to be present while the phenomenal states are absent or different. That is, materialism requires that entailments $P \supset Q$ be necessary, where P is the complete physical truth about the world and Q is an arbitrary phenomenal truth (Chalmers, 2010, pp. 115, 116).

This makes two very distinct kinds of Type-B physicalist. One is committed to an *identity* thesis and one is committed to a *non-identity* thesis.

Instead of the metaphysical relation of identity, one can naturally appeal to realization, constitution, or grounding metaphysical relations at this point. Applying ONTOLOGICAL AND CAUSAL IRREDUCIBILITY to phenomenal properties (or objects), we can say that the following is an option of the physicalist:

MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY: Some mental entities and features are (I) *not identical* to the configurations of fundamental microphysical objects and features

they depend on and are (II) distinctively causally efficacious with respect to the configurations of fundamental microphysical objects and features they depend on (Paraphrased from Wilson, 2021, pp. 1-2).

Just like ONTOLOGICAL AND CAUSAL IRREDUCIBILITY, MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY takes mental properties to genuine metaphysical properties, distinct from the fundamental microphysical properties. However, they are physicalistically acceptable because they are supposed to be distinctly causally efficacious in some sense consistent with NOMOLOGICAL CLOSURE and GLOBAL SUPERVENIENCE.

Conclusion

In this chapter, I have tried to explain three concepts: free will, determinism, and physicalism (of both the reductive and non-reductive varieties). I have suggested that the clearest path forward for the physicalist inclined to believe in free will is to be a non-reductive physicalist, because it is *prima facie* compatible with free will. In the next chapter, I will begin the process to test that incompatibility. However, to do this, I need to introduce the reader to the (traditional) consequence argument, which is not an argument about physicalism and free will primarily, but about determinism and free will.

CHAPTER TWO. THE P-MODAL CONSEQUENCE ARGUMENT

Introduction

In the last chapter, I defined free will and determinism (and physicalism). The consequence argument (Van Inwagen, 1983) is a famous argument for the incompatibility of free will and determinism. In this chapter, I shall explain the *modal* consequence argument as articulated by Huemer (2000), Speak (2011), and Pruss (2013). Following Pruss (2013), I will defend the validity of this argument against the most famous objection to the consequence argument: Agglomeration (McKay & Johnson, 1996). I will also consider an objection to this argument from Michael Mckenna and Derk Pereboom (Pereboom & McKenna, 2016). In the next chapter, I am going to take the basic logical structure of the modal consequence argument and articulate a new version of the argument. As such, the main goal of this chapter is to explain the original modal consequence argument and defend its validity.

I will begin by overviewing the basic intuitive motivation for the argument (2.1). Then, I will lay out the predominant formal modal interpretation of the consequence argument in 2.2 and explain why McKay and Johnson's coin-flipping agglomeration objection (1996) is problematic for the argument. In 2.3, I will explain Pruss' updated version of the modal consequence argument. The upshot of Pruss' version-particularly Pruss' updated Beta Rule and N-Operator- is that it can help us to overcome the problem(s) shown by McKay and Johnson (1996). Then, in 2.4, I will consider another objection (related to McKay and Johnson's objection) from Pereboom and Mckenna (2016). In 2.5, I will provide two responses to Pereboom and Mckenna in order to show that this objection is not successful. Having shown that both McKay and Johnson's objection and Pereboom

and McKenna's objection fails, I will have shown that the basic structure of Pruss' consequence argument is logically valid and worth further consideration.

2.1 The Consequence Argument

The consequence argument is one of the most discussed and famous arguments in the free will literature. In this section, I will offer some general outlines of the argument and explain why it is so intuitive. I will begin by discussing some common-sense examples from Michael Huemer (2021) that are similar to the reasoning in the consequence argument and then present Van Inwagen's informal representation of the consequence argument (1983)

Most discussions of the consequence argument utilize some sort of transitivity case to help show the intuitiveness of the argument. Michael Huemer provides the following (relatable) hypothetical situation:

A student asks me, near the end of the semester, "Hey Prof, how can I get an A in your class?" I reply: "In order to get an A in my class, you would have to have scored at least 87.5% on the first four tests. But in fact, you scored only 73%." The student responds, "So you're saying I can't get an A." (Huemer, 2021, 188).

The student's response is clearly reasonable. This student has no ability *in the present* to earn an A in the class because she did not receive adequate grades on her prior work (which strictly entails her final grade). Of course, in this scenario, the student had an ability to change her grades *in the past*. But the student's response is reasonable in this context.

Here is another hypothetical case from Huemer (2004, 414):

You are a doctor in the emergency room of a hospital, where a heart attack victim has just been brought in. You know that in order for such a patient to be revived, CPR must be administered within three minutes of the time of the heart attack. You also know that for this patient, four minutes have already elapsed, during which the patient did not receive CPR. You conclude that, at this point, you cannot revive the patient. (2004, 414)

Just like the student in first scenario, the doctor makes a reasonable inference. He was not with the patient four minutes prior to this moment. Therefore, he could not administer CPR four minutes ago. Furthermore, unlike the student in the first scenario, the doctor plausibly has never had an ability to change the fate of this patient *because* there is something in the past entailing that the patient will die which the doctor has no ability to change. This is the sort of intuitive inferential reasoning is at the heart of what is called the consequence argument.

The consequence argument is famously expressed in Peter Van Inwagen's *Essay on Free Will* (1983) in the following informal passage:

If determinism is true, then our acts are the consequence of laws of nature and events in the remote past. But it's 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 (Van Inwagen, 1983, p. 56).

Just like the doctor case just mentioned, for many people this is a very intuitively plausibly argument. I cannot make a difference toward what a T-Rex did on a given day in the Jurassic age, nor can I make any difference toward whether or not Caesar crossed the Rubicon. Additionally, it is extremely plausibly that the laws of nature are not up to me in any significant way. It is not as if, once I adequately understand Einstein's special relatively theory or the Schrodinger equation, I

can simply alter them to fit my current desires.³⁶ Perhaps a superhero, or time-traveler, or divine-miracle worker could change the laws of nature, but an ordinary person cannot change or alter the laws of nature regardless of how much one wishes she could.

In the doctor-CPR case, the fact that the patient *did not* receive CPR within four-minutes ultimately determined the patient's fate. In that world, the doctor cannot do anything to change the patient's fate. In the case of determinism, the completed state of the world in the remote past in conjunction with the laws of nature ultimately determine what I do in the present. If determinism is true, there is a very real sense in which I cannot change what I do in the present (on account of my inability to change the past or the laws). Therefore, I cannot have free will in a deterministic universe.

However, the passage from Van Inwagen is just an informal passage. An intuitive passage, but an informal one. There have been multiple ways to interpret this passage formally into a standard-form argument. Overlooking every interpretation of the consequence argument is not necessary here because I am defending a particular version of the argument (and not every version). Next, I shall discuss a particular modal interpretation of the consequence argument.

³⁶ I have thought about the question: on what occasion would I actually *want* to change the laws of physics? Perhaps if I was a poor athlete I would want to jump for enormous lengths by summoning quantum tunnels to shift mass around. Perhaps time-traveling would be something I would want to do under some circumstance. But, all of these seem more undesirable when I consider the possible consequences of such actions. The only real occasions I can think of where I would want to perform a miracle (defined as altering the laws of physics) is in cases of healing sick/hurt loved ones.

2.2 The Modal Consequence Argument

Following Pruss (2013), Speak (2011), and Huemer (2000), I am going to express an inherently modal formulation of the consequence argument next.³⁷ In order to run the consequence argument modally, we need to define a few principles and operators. I shall overview these principles before laying out the argument.

The consequence argument uses a symbolic N-Operator (e.g., Np) to express “p and no one has, or ever had, any choice about whether p” (Van Inwagen, 1989: 404). Mathematical propositions are helpful examples of when this N-designation could be applicable. The proposition “ $2+2=4$ ” (p) is a true proposition. It is also a necessary proposition. And, as Van Inwagen notes, it is difficult to conceive of anyone having the ability to change a necessary proposition; indeed, he says, “[h]ardly anyone besides Descartes has been willing to concede such a capacity even to God. No one, so far as I know, has ever suggested that human beings could have a choice about what is necessarily true” (Van Inwagen, 1983, p. 96). Put another way, at this moment, there is a range of available actions that I can perform (e.g., drinking coffee, swimming with my kids, or taking my wife on a date). However, nowhere in this set of possible actions is there a “change “ $2+2=4$ ””

³⁷ In Van Inwagen’s original *Essay on Free Will* (1983), he presented three formal versions of the argument. The first version of the argument is in a counterfactual form (Van Inwagen, 1975, 1983: 78-93). The first version is relevant to a prominent family of objections to the consequence argument, inspired by David Lewis (1981), but argued, additionally, by Kadri Vihvelin (2013, pp. 155-167), Carolina Sartorio (Sartorio & Kane, 2022, pp. 105-108), Cian Dorr (2016), and others. In § 5.2.5, I will discuss one version of this objection from John Maier (2022). The second version is called an “extension” argument by Campbell (2017, pp. 158-160), and does not employ any kind of principle that claims certain events are ‘not up to us.’ Instead, in this version, Van Inwagen focuses on whether or not we can genuinely add to the actual past (Van Inwagen, 1983, p. 167). This argument, as I understand it, would not be applicable, structurally, for my further argument in chapter three, so I will not address it further here. Among the three versions, the third version is the best on its own merits. In this chapter, I shall focus on Van Inwagen’s third version of the consequence argument (1989; 2000; 2008).

option. Other intuitive candidates for the application of N-operator would be propositions like “Saturn has seven rings,” or “penguins are birds.” Whether or not penguins are birds or Saturn has seven rings is not something ordinary human persons have ever been given a choice about in any meaningful sense. Thus, they are intuitively plausible candidates for the application of this N-operator. (Traditionally, philosophers have labelled principles utilizing the N-Operator as *fixity principles*. So if “penguins are birds is true and no one has, or ever had, any choice about whether p is true” then we can refer to this as (for example) the Fixity of the Penguins-As-Birds principle.)

In the consequence argument, the two typical fixity principles that use the N-Operator are:

FIXITY OF THE PAST: A proposition expressing the completed state of the world in the remote past (P) is true and no one has, or ever had, any choice about whether P is true (i.e., NP is true).

FIXITY OF THE LAWS: A proposition expressing the completed laws of nature (L) is true and no one has, or ever had, any choice about whether L is true (i.e., NL is true).

FIXITY OF THE PAST (which can be referred to as simply NP) is a principle concerning some completed state of the world in the remote past. Now, this is usually thought to describe the state of the world in the *pre-human past*. So, we can imagine a particular time *t* in, say, the Jurassic period when a T-rex is fighting a triceratops and no human persons exist. (Although, it really can refer to any time in the remote past.) Here, it is quite intuitive to suggest that no human person has any choice whatever over what went on at *t*. Why? Because no one was there, or can be there assuming time-travel is impossible. FIXITY OF THE LAWS is simply suggesting that the (completed) laws of nature are what they are and we cannot change them. It is true that we do not have a

completed physics yet, but whatever it is going to look like, no one seriously thinks have the power to genuinely change a law of nature.³⁸ These are the two fixity principles in the modal consequence argument. Next, I need to discuss two key modal rules for the modal consequence argument.

Additionally, there are two original modal rules that are usually used in the modal versions of the consequence argument. The two rules both assume the N-operator just discussed, and they can be stated as follows:

ALPHA: $\Box p \vdash Np$

BETA: $Np, N(p \rightarrow q) \vdash Nq$

Rule Alpha is stating that if p is necessary, then no one has any choice about it. Rule Beta is a transfer rule stating that if no one has any choice about p, and no one has any choice about p entailing q, then no one has any choice about q.³⁹

In order to express the modal consequence argument, the last thing I need to discuss is a modal definition of determinism, or, rather, a modal description of a freely willed decision *in a deterministic universe*. (Of course, determinism is a much broader thesis than just what I am about to offer as a principle). Let α be the proposition that “Hadassah decides to attend the University of

³⁸ As I will mention later, sometimes these are combined together (e.g., $N(P \wedge L)$), but for now I will consider them separately.

³⁹ I will spend more time in a moment discussing these rules in depth as these rules (especially BETA) are controversial. However laying out the modal rules in this simple fashion will be enough to introduce what the modal consequence argument looks like formally.

Birmingham at t .” This is a proposition concerning a plausibly freely willed action by a human person in the present (so assume t is a time in the present). Additionally, let L be a proposition stating a completed set of the laws of nature and P be a proposition stating the completed state of the world at some time in the pre-human past (as previously discussed). Given determinism, we can infer the following is true:

INTENTIONAL ACTION GIVEN DETERMINISM: Necessarily, if the conjunction of P and L is true, then α is true (or, $(\Box((P \wedge L) \rightarrow \alpha))$).

This principle (i.e., INTENTIONAL ACTION GIVEN DETERMINISM) is, again, supposed to be a modal articulation what how any decision or action would be related to P and L in a deterministic world.

Using the two modals rules, Pruss (2013, 430-431) puts the modal consequence argument in a standard Fitch-style proof using the following modal rules from system T: Distribution ($\Box(A \rightarrow B) \vdash (\Box A \rightarrow \Box B)$); Necessitation of Tautology (if p is a tautology, then $\vdash \Box p$); Rule T ($\Box p \vdash p$).

From these rules we can put the following argument I will refer to as the MODAL CONSEQUENCE ARGUMENT (which is slightly paraphrasing Pruss’ argument for the purpose of keeping certain symbols consistent, from 2013, 431):

- | | |
|--|--------------------------------|
| (1) $\Box((P \wedge L) \rightarrow \alpha)$ | Assumption from determinism |
| (2) NL | Assumption from FIXITY OF LAWS |
| (3) NP | Assumption from FIXITY OF PAST |
| (4) $\Box((P \wedge L) \rightarrow \alpha) \rightarrow (P \rightarrow (L \rightarrow \alpha))$ | Nec Taut |
| (5) $\Box((P \wedge L) \rightarrow \alpha) \rightarrow \Box(P \rightarrow (L \rightarrow \alpha))$ | 4, Distribution |
| (6) $\Box(P \rightarrow (L \rightarrow \alpha))$ | 1,5, Elimination |
| (7) $N(P \rightarrow (L \rightarrow \alpha))$ | 6, ALPHA |

(8) $N(L \rightarrow \alpha)$	2,7 BETA
(9) $N\alpha$	3,8, BETA

Premises (1)-(3) is expressing INTENTIONAL ACTION GIVEN DETERMINISM and the two intuitive fixity rules separately. Premise (4) is introducing a necessary tautology into the argument because $(P \wedge L) \rightarrow \alpha$ and $(P \rightarrow (L \rightarrow \alpha))$ are going to be equivalent. Once, we have (1)-(4), we can apply the distribution rule to infer (5) is true, and (1) and (5) infer (6). Steps (4)-(6) are odd to read, but they are logically valid moves. (These are also not really controversial, some articulations of the modal consequence argument just skip over from (3) directly to (6) and say that it is from logical entailment). Premise (7) is inferred from (6) and ALPHA. Then BETA will ultimately entail (8) and (9).

However, some readers may wonder why this argument is important for free will. After all, there is no mention of free will in this argument. This is because there is an implicit assumption in the argument: if $N\alpha$ is true, then necessarily that no one has free will with regards α . Put another way: if no one has a choice about a (including Hadassah), then no one has free will with regards to a (including Hadassah). If $N\alpha$ stands for “ α is true and no one has, or ever had, any choice about whether α ” then it stands to reason that Hadassah did not have any choice about a . That claim *prima facie* is going to be hard to make coherent with any theory of free will. Once this assumption is made, one could infer from (9) that no one, not even Hadassah, has free will with regards to her choice to attend Birmingham. However, this is problematic because this reasoning can be applied to all human decisions. Thus no one, ever has free will. Because this argument is made in a standard proof method the conclusion-assuming this implicit assumption- is not *technically* that no one has free will; rather, the conclusion is that necessarily if determinism is true, then no one can have free will. To put it in possible world language, there are no possible worlds where

determinism is true *and* there are persons or creatures with free will. This is the modal consequence argument.

What is particularly interesting about the modal consequence argument (presented above) is that there are really only four good candidates for criticism. Very few philosophers would criticize (1) or (4)-(7). However, objectors commonly disagree with NP (3) or NL (2), or BETA (the inference rule that entails (8) and (9) from previous premises), or the implicit assumption that $N\alpha$ necessarily that no one has free will with regards α . For the moment, let us assume the implicit assumption that $N\alpha$ necessarily that no one has free will with regards α is a true assumption. NP and NL are obviously *prima facie* intuitive principles (i.e., no one really thinks they have a choice over the laws of nature or the remote past).⁴⁰ So, the natural candidate for criticism is BETA. And, there is one small (or, perhaps not so small) problem for BETA: the fallacy of agglomeration, which I shall turn to next.

2.2.1 Agglomeration

I noted before that there are four plausible candidates to criticize in the modal consequence argument. In this section, I shall discuss what most critics of the modal consequence argument think is the most fatal objection. The agglomeration problem is a problem for BETA, which was defined as $Np, N(p \rightarrow q) \vdash Nq$.

⁴⁰ Which does not mean they are necessarily true.

Van Inwagen explains that his confidence in BETA is justified primarily by lacking awareness of any counterexamples to it:

I must confess that my belief in the validity of beta has only two sources, one incommunicable and the other inconclusive. The former source is what philosophers are pleased to call 'intuition'.... The latter source is the fact that I can think of no instances of beta that have, or could possibly have, true premises and a false conclusion (Van Inwagen 1983, pp. 97-98).

There is nothing wrong with positing a logical rule because it is highly intuitive and has no counterexamples. However, if the rule is not derived from other clearly established rules, then it is obviously a plausible candidate to be attacked. This admission from Van Inwagen led many critics to attempt to find a counterexample to BETA in order to disprove the modal version of the consequence argument.

This critical endeavor culminated chiefly in the following sort of counterexample (McKay & Johnson, 1996): Imagine that I have an indeterministic coin and think about flipping it, but ultimately decide to put it back in my pocket (perhaps I am shy about my coin-flipping skills). Let p be the proposition the coin does not land on tails and q be the proposition that the coin does not land on heads. One would likely accept that Np because, as McKay and Johnson (1996, p. 115) note, “no one can choose to make the coin fall heads” (after all its an indeterministic coin!). And the same reasoning would work for Nq . However, $N(p \wedge q)$ is false because I have a decision in front of me to flip the indeterministic coin, which would make *either p or q false*. The conjunction (i.e., $p \wedge q$) is supposed to be true in the circumstance where you do not flip the coin because it does not land on either. However, if either p or q is false, then the conjunction of p and q is false. As Johnson and McKay say, we “could have chosen to make $[p \wedge q]$ false by choosing to

toss the coin” (ibid). This is an example of agglomeration, which, in this example, shows that one could have true premises and a false conclusion while using BETA

Agglomeration is relevant for the consequence argument because, as BETA and ALPHA are defined above, the following proof—where we can input the proposition p as “the coin does not land on tails” and q be the proposition that “the coin does not land on heads” — seems valid (Pruss, 2013, p 434):

- | | |
|---|----------------|
| (1) Np | |
| (2) Nq | |
| (3) $\Box[p \rightarrow (q \rightarrow (p \wedge q))]$ | Nec Taut |
| (4) $N[p \rightarrow (q \rightarrow (p \wedge q))]$ | 3, Rule Alpha |
| (5) $N (q \rightarrow (p \wedge q))$ | 1,4, Rule Beta |
| (6) $N (p \wedge q)$ | 2,5, Rule Beta |

This proof is valid by BETA and ALPHA. However, this proof *should not be valid* because it is agglomerative and the conclusion is false. However, in this proof the only step or inference rule one could reasonably reject is at least one of either BETA or ALPHA, and Rule Alpha seems far more plausible than BETA and so, therefore, BETA should not be considered a valid modal inference rule.

What I have just shown is that examples like the indeterministic coin toss are agglomerative examples where, using BETA, one can create an argument with true premises and a false conclusion. Thus, something *must* be wrong with the agglomerative argument. The best candidate is BETA. This matters for the modal consequence argument because, if BETA is not a valid modal inference rule, then the modal consequence argument is not valid (and, thereby, not sound).

The agglomerative argument is, by far, the most discussed criticism of the consequence argument, and the most plausible. Van Inwagen even admits that McKay and Johnson disproved BETA (or “ β ” as Van Inwagen refers to it below):

And Thomas McKay and David Johnson have shown that β is invalid...The reason I mistakenly supposed that β was valid was this. I mistakenly supposed that the only way in which it could be that one had no choice about the truth-value of a proposition would be for the truth-value of that proposition to be in some way so firmly “fixed” that one was unable to change it. I did not see that that there is another way for one to have no choice about the truth-value of a proposition: for that truth-value to be a mere matter of chance.
(Van Inwagen, 2015, 18-19)

Choosing to argue against the modal consequence argument by using the agglomerative argument is dialectically advantageous because, compared to the NP or NL (which are pre-theoretically intuitive principles), there is no real pre-theoretical intuitiveness behind Beta. A critic does not need to change anything metaphysically radical in order to reject the modal consequence argument if all one needs to do is reject Beta (juxtaposed this to the critic who theoretically rejects NP, this critic must argue that ordinary persons *have a choice over the remote past*, a seemingly radical metaphysical thesis).

Next, I am going to explain why proponents of the consequence argument have not given up all hope. They have not given up hope because they believe they can defend Rule Beta (or a version of it) against these sorts of fallacious agglomerative examples.

2.3 Pruss' Modal Consequence Argument

In response to the previous criticism of BETA, the proponents of the modal consequence argument have an obvious option: to revise BETA. However, this is theoretically difficult because, if the only reason one accepts a particular articulation of Beta is to avoid the coin-toss example, this new revised Beta would obviously be *ad hoc*. Instead, what the proponents need to do is to find a version of Beta that has some justifiable reason for its existence. If a revised form of Beta can be discovered the modal consequence argument would potentially be immune to such a McKay and Johnson-styled objection.

2.3.1 Beta-Best

The original candidate for the new Beta was first introduced by Widerker (1987) and Finch and Warfield (1998). It is the following:

$$\text{BETA-BEST: } Nq, \Box(q \rightarrow p) \vdash Np$$

The essential difference between this BETA-BEST and BETA is that BETA used the N-operator outside of the conditional claim in the second premise, while Beta-best uses a necessity operator in the claim in the second premise. One benefit of BETA-BEST is that it allows one to potentially dispense with ALPHA altogether. If one writes the fixity principles in a conjoined conjunction, then one can write the modal consequence argument in the following way (again, using P as the Past, L as the laws, and α as a proposition expressing Hadassah deciding to attend Birmingham in the present):

- | | |
|---|---------------------------------|
| (1) $\Box((P \wedge L) \rightarrow \alpha)$ | Modal Definition of Determinism |
| (2) $N(P \wedge L)$ | Fixity Principle |
| (3) $N\alpha$ | 1,2, BETA-BEST |

In this version of the modal consequence argument-assuming we can conjoin the fixity principles-there is a much simpler inference from (1) to (3). Given BETA-BEST, there is no need to employ ALPHA.

Additionally, agglomeration is not a problem given BETA-BEST. Consider the following attempt to prove fallacious agglomeration:

- | | |
|--|----------------|
| (1) Np | |
| (2) Nq | |
| (3) $\Box[p \rightarrow (q \rightarrow (p \wedge q))]$ | Nec Taut |
| (4) $N(q \rightarrow (p \wedge q))$ | 1,3, Rule Beta |
| (5) $N(p \wedge q)$ | ? |

It is hard to see how one would argue to $N(p \wedge q)$ (5) from (1)-(4). One could do that with BETA (the original rule), but BETA-BEST uses the necessity operator which limits the inferences one can make from premises (1)-(3).

I have just mentioned two benefits of BETA-BEST, but there is still one clear problem: it appears *ad hoc*. The only reason it seems one would need to accept BETA-BEST is that the rule would help us to avoid the fallacious agglomerative example from McKay and Johnson (1996).

Pruss (2013), however, offers a serious non-arbitrary motivation for BETA-BEST. His argument for BETA-BEST is based on a modal rule called “Weakening”. In Lewisian semantics, Weakening is a rule that states the following (where $\Box \rightarrow$ stands for counterfactual sufficiency):

Weakening: If p is counterfactually sufficient for q , and necessarily q implies r , then we can infer that p is counterfactually sufficient for r (Or, $p \Box \rightarrow q$, $\Box(q \rightarrow r)$, therefore $p \Box \rightarrow r$).

For example, imagine I am playing a card game and drawing a spade would guarantee me victory in the game. Furthermore, if I win this game then I will necessarily be the winningest poker player in the history of the casino. If this is true, then it will be true that if I draw a spade, I would be the winningest poker player in the history of the casino. Pruss succinctly explains this rule:

In Lewis' semantics, weakening is easy to show. If p is impossible, then $p \Box \rightarrow r$ holds trivially. If p is possible, then $p \Box \rightarrow q$ holds if and only if there is a world w_1 in which p and q hold and which is closer to the actual world than any world at which p and $\sim q$ hold. Let w_1 be such a world. Since q entails r , we must also have r holding at w_1 . Let w_2 now be any world at which p and $\sim r$ hold. Because q entails r , this is a world at which p and $\sim q$ hold and hence this world is further from the actual world than w_1 , is, by choice of w_1 . Thus, $p \Box \rightarrow r$ holds (Pruss, 2013, 322).

From Weakening, Pruss wishes to argue that BETA-BEST can be proven as a valid modal inference rule.

However, to prove BETA-BEST from Weakening, Pruss needs a way of talking about the N-operator in a modally more sophisticated way. Following Huemer (2000), Pruss reads Nr as “saying that there is nothing that anyone can (ever) do that would falsify r ” (Pruss, 2013, p. 433), or, modally as:

N-Operator: Nz if and only if $z \wedge \neg \exists x \exists \alpha [\text{Can}(x, \alpha) \wedge (\text{Does}(x, \alpha) \Box \rightarrow \neg z)]$

Pruss (following Huemer, 2000, 529) definition of N is suggesting that, in order for Nz to be the case, z must be true and there cannot exist any person who would have an ability α such that if she were to exercise α , z would be false. Here, $\Box \rightarrow$ is a subjunctive would-conditional, and x ranges over human persons and α ranges over all action or ability types (ibid). Pruss notes that it is inevitably going to be controversial as to the best way to analyze $\text{Can}(x, \alpha)$, but “it will turn out that we will not need controversial analysis- our proof works regardless of how we analyze it.”⁴¹

To illustrate this imagine there was a vial of magical potion. On the label of this vial, the label read that this potion would cure your disease, but a possible side-effect was that it could turn you invisible. The label assures you though that this is a one out a million chance. On this definition of $N(x)$, we can ask whether or not we could apply N to the proposition “no human person is ever invisible” (x). In this scenario there is an available action you could perform (i.e., you could drink the vial) that has a possibility of falsifying x . However, your drinking the vial is not counterfactually sufficient (in a would-sense) for $\neg x$ being true. This is because although there is a possible world where you drinking the vial and x false, there are far (far) more possible worlds where you drink the vial and x is true, and these worlds are more similar (and therefore closer) to the actual world. Therefore, although there are different criteria for evaluating counterfactual sufficiency, on most views, drinking the potion will not be sufficient for turning invisible because the world where $\neg x$ is true is too far away from the actual world.⁴²

⁴¹ This is different than the intuitive “no choice about” version in the original modal consequence argument.

⁴² If it was a “might-conditional” perhaps it would not turn out the same.

Using this new N-Operator and BETA-BEST, Pruss (2013, p. 434) uses the same rules as he did in the original modal consequence argument, with the exception of the addition of Weakening and using z and q for any two propositions:

- (1) Nz
- (2) $\Box (z \rightarrow q)$
- (3) $z \wedge \neg \exists x \exists \alpha [\text{Can}(x, \alpha) \wedge (\text{Does}(x, \alpha) \Box \rightarrow \neg z)]$ Def, 1
- (4) $\neg \exists x \exists \alpha [\text{Can}(x, \alpha) \wedge (\text{Does}(x, \alpha) \Box \rightarrow \neg z)]$ $\wedge$ -Elim, 3
- (5) $\forall x \forall \alpha \neg [\text{Can}(x, \alpha) \wedge (\text{Does}(x, \alpha) \Box \rightarrow \neg z)]$ De Morgan, 4
- (6) $ba \mid \neg [\text{Can}(b, \alpha) \wedge (\text{Does}(b, \alpha) \Box \rightarrow \neg z)]$ $\forall$ -elim, 5
- (7) $\mid \text{Can}(b, \alpha) \rightarrow \neg (\text{Does}(b, \alpha) \Box \rightarrow \neg z)]$ Taut con, 6
- (8) $\mid \Box [z \rightarrow q] \rightarrow (\neg q \rightarrow \neg z)]$ Nec Taut
- (9) $\mid \Box [z \rightarrow q] \rightarrow \Box (\neg q \rightarrow \neg z)]$ Distr, 8
- (10) $\mid \Box (\neg q \rightarrow \neg z)$ $\rightarrow$ -Elim, 2, 9
- (11) $\mid \text{Can}(b, a)$
- (12) $\mid \mid \text{Does}(b, a) \Box \rightarrow \neg q)$
- (13) $\mid \mid \text{Does}(b, a) \Box \rightarrow \neg z)$ Weaken, 10, 12
- (14) $\mid \mid \neg (\text{Does}(b, a) \Box \rightarrow \neg z)$ Elim, 7, 11
- (15) $\mid \neg (\text{Does}(b, a) \Box \rightarrow \neg q)$ Reductio 12-14
- (16) $\mid \text{Can}(b, \alpha) \rightarrow \neg (\text{Does}(b, \alpha) \Box \rightarrow \neg q)$ $\rightarrow$ -Intro 11-15
- (17) $\mid \neg [\text{Can}(b, \alpha) \rightarrow \neg (\text{Does}(b, \alpha) \Box \rightarrow \neg q)]$ Taut con, 16
- (18) $\forall x \forall \alpha \neg [\text{Can}(x, \alpha) \wedge (\text{Does}(x, \alpha) \Box \rightarrow \neg q)]$ $\forall$ -intro, 17
- (19) $\neg \exists x \exists \alpha [\text{Can}(x, \alpha) \wedge (\text{Does}(x, \alpha) \Box \rightarrow \neg q)]$ De Morgan, 18
- (20) $z \rightarrow q$ T, 2

(21)	z	$\wedge$ -Elim, 3
(22)	q	$\rightarrow$ -Elim, 20, 21
(23)	$q \wedge \neg \exists x \exists \alpha [\text{Can}(x, \alpha) \wedge (\text{Does}(x, \alpha) \Box \rightarrow \neg q)]$	$\wedge$ -intro 19,22
(24)	Nq	Def, 23

This proof is valid. Thus, one cannot suggest that BETA-BEST is only justified because of an *ad hoc* motivation. This proof is not really controversial in that the only way one could deny the conclusion is to deny weakening.

At this point, one might ask whether one could still come up with a counterexample to BETA-BEST (perhaps based not on agglomeration, but rather some other fallacious result). Pruss notes that, if that did occur, “it would be a substantive philosophical achievement, since it would also be a counterexample to weakening and hence to Lewis's semantics for counterfactuals, as well as to various modified versions of that semantics on which weakening still holds” (Pruss, 2013, p. 434).

43

⁴³ There are some new potential counterexamples in the recent literature, but they mainly have to do with various interpretations of the modal actuality operator. Here is one such example:

Now let ϕ be the statement, say, that Nina raises her hand. Then someone has a choice about whether ϕ is true, namely, Nina herself. Moreover, if Nina actually exists, she has a choice about whether ϕ is true at the actual world. But from this fact it is reasonable to conclude that she also has a choice about whether the actualized statement $@\phi$ is true at the actual world, even though $@\phi$ and ϕ have different modal profiles: the former is necessary if true, whereas the latter is only contingent (Lambert and Merluzzi, 2020, 452)

Here, Lambert and Merluzzi are suggesting that Nina has a choice about $@\phi$ and $@\phi$ is a necessary truth. This is a potential counterexample to BETA-BEST because one argues $Nq, \Box(q \rightarrow @\phi) \vdash N@\phi$, where q is any proposition, $N@\phi$ will logically follow (because the consequent in the conditional will be true because it is necessary). However, they claim, $N@\phi$ is false. To explain this a little further, note that Lambert and Merluzzi define the actuality operator

In this section, I have shown that Pruss’ proof for BETA-BEST from Weakening is valid and that BETA-BEST is also immune to the McKay and Johnson agglomeration counterexample. Additionally, I have introduced Pruss’ new interpretation of the N-operator. In the next section, I am going to formally introduce Pruss’ modal consequence argument with BETA-BEST and the updated N-operator.

2.3.2 The P-Modal Consequence Argument

So, where does this leave the argument? Well, Pruss’ proof for BETA-BEST from Weakening provides strong motivation for accepting BETA-BEST, which allows the consequence argument to be expressed simply as (again, using the updated N-operator, and P as a completed state in the Past, L as the completed fundamental laws of nature, and α as a proposition expressing Hadassah deciding to attend Birmingham in the present):

- | | |
|--|---------------------------------|
| (1) $\Box (P \wedge L) \rightarrow \alpha$ | Modal Definition of Determinism |
| (2) $N(P \wedge L)$ | Fixity Principle |

(@) in the following way: “@ ϕ is actually true at a possible world w if and only if ϕ is true at the actual world of the model, say, w^* ” (2020, 451). They additionally say, “Why do we believe that Nina has a choice regarding the truth of @ ϕ at the actual world? Well, we believe this because the truth of both @ ϕ and ϕ obtain at the actual world for the very same reasons, that is, the very same facts, grounds, states of affairs, truthmakers or what have you, that make @ ϕ true at the actual world are the ones that make ϕ true at the actual world” (Lambert and Merluzzi, 2021a, 452). This is a really interesting modal puzzle to work out, but I take it that this is chiefly a problem for how to work out modal indexicals. I don’t see how, contra these authors, Nina would have it in her power to make @ ϕ false if @ ϕ practically means “ ϕ is true *at such-and-such possible world*.” No one has power to change which propositions are true in which possible worlds. There might a sense in which we have the power to determine which possible world is actual. But, overall, I take it that the authors are assuming a particular modal system that ends up with this logical result. But this objection is something I would like to consider further. C.f., Turner and Capes (2014).

(3) $N\alpha$

1,2, BETA-BEST

This argument suggests that, if determinism is true, Hadassah has no ability to falsify α . She cannot prevent or change her decision to attend Birmingham. Of course, we could also run the argument where a stands for the completed state of the world in the present, which would, then, subsume all of Hadassah's actions, thus Hadassah cannot do anything other than what she actually does.

Recall, in Chapter One (§1.1.1), I defined free will in such a way that the following necessary condition is true:

ALTERNATIVE-POSSIBILITIES: A creature C is free in ϕ -ing *only if* C , when choosing to ϕ , could decide to $\neg \phi$.

From this necessary condition of free will it is clear that $N\alpha$ entails that Hadassah cannot have free will because she cannot do anything other than what she actually does (i.e., she cannot falsify α). Thus, more formally, the argument adds the following premises (although some proponents fail to explicitly include this) where Fx can stand for the designation “someone has free will with regards to x ”:

(4) $N\alpha \rightarrow \neg F\alpha$

(Assumption from Definition of FREE WILL)

(5) $\neg F\alpha$

(4,3, MP)

Thus, if we read the argument as a conditional proof, then the argument would conclude that *necessarily*, if determinism is true, no one can have free will:

(6) $\Box (\Box (P \wedge L) \rightarrow \alpha) \rightarrow \neg F\alpha$ (CP)

The modal consequence argument therefore is a very powerful argument for this particular Incompatibilist thesis, because the premises seem very intuitive (no one naturally thinks they can change the past or the laws) and the modal inference rules seem valid too.

However, while this argument has convinced many people of its conclusion, several critics have made different kinds of objections to it. Next, I turn to another objection to try to illuminate further aspects of the argument that might not be obvious at this point.

2.4 Mckenna and Pereboom's Objection

There are a few notable objections or lines of criticism relevant for the P-MODAL CONSEQUENCE ARGUMENT.⁴⁴ One objection that I will point out here comes from Pereboom and Mckenna (2016, pp. 82-84, 98), and is related to the objection to McKay and Johnson's.⁴⁵ Pereboom and McKenna

⁴⁴ Joseph Campbell (2007; 2010), for example, has written a novel criticism of the argument. His criticism asks us to imagine God creates a deterministic universe (relevantly similar to our own), and, in addition to this world, creates an adult male Adam *ex nihilo* with the world (Campbell, 2007). Thus, there is no past for Adam. So theoretically, assuming the P-Modal Consequence Argument is sound, Adam could be free still in a deterministic universe. In of itself, this does not seem to be a reason to believe the CA fails. However Campbell presses the hypothetical, what if, then, God creates Eve. However, Eve *would* have a past, and, presumably, lack free will. He then goes to argue that "it is hard to see how Adam and Eve differ in any important respect" (Campbell 2007, 110). Daniel Speak (2011) calls this "The Eve Challenge." It is hard to see how Eve and Adam differ in terms of their intrinsic metaphysical properties, yet the Modal Consequence Argument would not obviously rule out the possibility of Adam having free will, yet, it seems, Adam and Eve should either both possibly have free will or not. This is an interesting case and might pose a problem, but there are two responses that I think are sufficient. Daniel Speak (2011) points out the consequence argument (generally) is "not committed to the claim that Adam can be free" in this *ex nihilo* Adam world, and in fact, it could be the case that "some other argument might show that Adam cannot be free" (126). This is one kind of response. Another is to point out that decisions and actions take a certain amount of time. The completed state of the world in the moment Adam entered existence (where this moment is a shorter time interval than it takes to make a decision) can serve where the remote past serves in the original argument.

⁴⁵ This is inspired from an objection originally from Michael Slote (1982). This is either the same objection or a very similar objection to Al Mele's (2006, p. 138) complaint about the consequence argument. Although, Mele's criticism is not quite as clear as the way McKenna and Pereboom present it. In Mele's (2022), he does offer a second criticism from the success of Frankfurt cases, which I discuss in Chapter Five (§ 5.3).

point out that the motivation for NP appears to be different than the motivation for NL, and this could be problematic.

Pereboom and McKenna (2016, p. 98) explain that if one uses a conjoined fixity principle (e.g., $N(P \wedge L)$)-as opposed to distinct fixity principles (e.g., NP and NL separately)-then one is implicitly assuming that one can combine fixity principles into one conjunctive principle. In other words, a conjoined fixity principle (e.g., $N(P \wedge L)$) is only viable if this sort of power necessity has the property of agglomerative. We can only derive it from NP and NL distinctly, which is obviously problematic for the reasons discussed with the McKay and Johnson (1996) example. They go on to say, “If [the conjoined principle is agglomerative], then a viable version of the Consequence Argument can only proceed by introducing distinct premises for power necessity over the laws and power necessity over the past” (Pereboom and McKenna, 2016, p. 96). So, if one believes the best objection to the consequence argument is regarding agglomerativity, then perhaps one can press that problem against the conjoined fixity principle. This is an objection worth consideration.

However, I do think one can make an effective response to this worry. The obvious response is to suggest that the motivation for $N(P \wedge L)$ may not be disjointed in the manner they suggest (i.e., it is not from different motivations for NL and NP respectively). In my strategy, I owe an explanation of why $N(P \wedge L)$ is plausibly true *jointly* rather than just *individually*. This strategy, in my mind, seems more likely to succeed than splitting up the fixity principles.

My strategy depends on how one views the nature of the laws of nature. (In other words, it depends on how one answers the question: what is a law of nature?) As I see this issue, the key to motivate $N(P \wedge L)$ jointly is to focus on why NL individually would be plausible in the first place. There are many theories regarding the nature of the physical laws, but they all agree that laws of nature correspond or are associated with certain physical regularities. However, we can divide these

theories in two broad camps. To see this divide clearly, consider this fun hypothetical scenario from Tyler Hildebrand (2023, p. 1),

Virtual Physicists: The departments of Physics and Computer Science have collaborated to produce a video game, *Virtual Physicists*. The objective is to explore a simulated environment, make observations, collaborate with other players, and together formulate a scientific theory that explains and predicts events in the virtual world. The game ends when the players discover the true final theory: the set of lawlike generalizations and boundary conditions in the program running the simulation.

Hildebrand (2023, p. 2), continues,

Your task in the game is to answer *scientific* questions about laws. (It’s not called “Virtual Philosophers”!) But suppose we ask why the virtual world contains regularities as opposed to irregularities. The answer is obvious: The regularities – namely, patterns on your screen – are *not* simply a rock-bottom stopping point for explanation, nor do they occur purely by chance; rather, they are *imposed* on the virtual environment by the program and ultimately by the programmers/developers.

According to one broad group of theories, the regularities in the world are analogous to the regularities in this video game. In other words, the regularities in our world are governed and imposed by the laws of nature. Hildebrand calls this family of views “Non-Humeanism”:

Non-Humeanism: The world is like a grand mosaic, but it’s *not* the case that its tiles could have been arranged any which way. Something imposes structure on it, analogously to the way in which the computer program *Virtual Physicists* imposes structure on its virtual environments. For example, some Non-Humeans posit a god as the enforcer of natural

laws, some treat laws as primitives, and there are other options besides. But whatever the details, all Non-Humeans accept some sort of basic, fundamental necessity: something that governs, produces, or somehow constrains patterns of events in nature

That is, physical laws are above physical phenomena and (metaphorically) govern the behavior of physical phenomena to make sure all physical phenomena behave properly. Laws of nature would be logically prior to the physical phenomena (i.e., the regularities) they govern. There are several (very) different theories under this category such as Divine Voluntarism (e.g., Swinburne, 2006), Primitivism (e.g., Maudlin, 2007), the Dretske-Tooley-Armstrong (DTA) View (e.g., Dretske, 1977; Tooley, 1977; Armstrong, 1983), and Dispositionalism (e.g., Mumford, 2004; Vetter, 2018).

According to a second broad group of theories, laws are descriptions or generalizations of the physical phenomena themselves. All electrons behave in such and such manner, and therefore it is a law that they behave in such and such manner. Therefore, the laws of nature are not on this view logically prior, but rather logically posterior to the cosmos. This is the view that Hildebrand calls Humeanism, which he defines as follows:

Humeanism: The world is like a grand mosaic whose tiles could have been arranged any which way. Nothing is ultimately responsible for its regularities. They are just a basic, fundamental feature of our world. Statements of laws, then, are merely descriptions of the most significant regularities that happen to occur. Thus, our world is *not* analogous to the worlds in *Virtual Physicists*.

There are different versions of contemporary Humeanism (about the Laws) today (e.g., Lewis, 1986; Loewer, 2009; Albert, 2016; Beebe, 2000).⁴⁶

What is important is that at least on some theories of the laws of nature, there is an important link between the laws and the completed state of the world at a given time. On some theories the laws of nature are dependent on the properties of objects (e.g., Humeanism, Dispositionalism). Recall, NL is the fixity principle claiming that no one can do anything that would falsify a proposition expressing the completed set of the laws of nature. In other words, no one can change the laws of nature. Based on several of the theories of the laws of nature, the counter-intuitiveness of the ability to change the laws is not rooted in an inability to falsify necessary truths, but rather in something like an inability to change the fundamental nature of physical objects. Sometimes we think it is counter-intuitive to change a law of nature because they are primitive or necessary truths and we cannot change necessary truths. However, if the laws (in whatever sense) are dependent on the

⁴⁶ Some readers may be aware that Mele and Beebe (2002) argue that Humeanism (about the Laws) is actually a path to disagree with NL in Modal Consequence Argument, and, hence, $N(P \wedge L)$. They argue that Humeanism is compatibilist in the following sense (Mele and Beebe, 2002, pp. 207-208): if one takes the completed state of the world in the past at t (P), on Humeanism, the properties and relations present in the world until t are not enough to give a complete statement of the laws of nature. In order to give a complete statement of the laws of nature, one has to have “facts about future regularities” (208). From the position of a kind of angelic observer of the universe *at the end of time*, one could look back overall the whole history of the world and discern L , and then figure out that P and L entail all other facts in the universe. Mele and Beebe argue that one can have a choice over or change L by virtue of L ’s dependence on us analogous to our choice over any future proposition concerning contingent actions (e.g., it may be true “I will eat eggs tomorrow” but they would argue it does not obviously deprive me freedom). L is dependent on the regularities concerning the whole of history, which contains our past, present, and future actions, after all. However, there is a dilemma here that is relevant: Assume P is the completed state of the world in the past at t , and let L^1 stand for the completed set of laws of nature that would hold for the world from the beginning of time until t . So L^1 would not stand for the completed laws of nature on a Humean view. Would P and L^1 entail the present? If the answer is yes, then the Modal Consequence Argument is not threatened because we would merely change the fixity principle to the appropriate proposition $N(P \wedge L^1)$. This is, I think, the right answer. However, suppose the Humean wants to deny that P and L^1 entail the present. If P and L^1 together do not entail the present, then there is going to be possible world with the exact same history and laws up until t with different futures.

physical phenomena itself, then the motivation for NL and for $N(P \wedge L)$ is actually quite similar. One cannot change the events in the past (assumption), and therefore one cannot change the fundamental descriptions of how the objects and features in those events behave (the laws) because they cannot change the events or states in the past (which contain the objects and features, which have certain behaviors and dispositions).⁴⁷

2.4.3 The Formal Argument Again

From this I would conclude that, as long as a theory of the nature of the completed set of fundamental physical laws are descriptive or (in some sense) dependent on the physical properties rather than necessitation, the motivation is actually quite connected between NP and NL. Thus, at least, for the consequence argument, there is no obvious need to disjoint them at all. We can just argue in the following way for the P-MODAL CONSEQUENCE ARGUMENT:

- | | |
|--|---|
| (1) $\Box (P \wedge L) \rightarrow \alpha$ | Modal Definition of Determinism |
| (2) $N(P \wedge L)$ | Fixity Principle |
| (3) $N\alpha$ | 1,2, Rule Beta |
| (4) $N\alpha \rightarrow \neg Fa$ | (Assumption from Definition of FREE WILL) |
| (5) $\neg Fa$ | (4,3, MP). |

⁴⁷ On some Non-Humean views (DTA), the laws are a kind of modal relation amongst universals, (Hildebrand, 2023, pp. 31-32). Laws are higher order states of affairs on this view. These would be relevant states of affairs for a complete description of the world in the remote past, so I think, like Humeanism and Dispositionalism, this would be fairly workable with my suggested response above. I grant that Primitivism and Divine Voluntarism seems different than the others in this regard. However, on Divine Voluntarism, there is a different worry than simply determinism. I suppose the primitivism can argue that $N(P \wedge L)$ is not as intuitive given their view without agglomeration. However, I still think the conjunctive claim is jointly intuitive to some degree on its own.

(6) $\Box (\Box (P \wedge L) \rightarrow \alpha) \rightarrow \neg F\alpha$ (CP, 2-6)

Or,

- (1) A proposition expressing the completed state of the world in the remote past (e.g., a time in the Jurassic period) in conjunction with a proposition expressing a completed set of the laws of nature necessarily entails the completed state of the world in the present (including some decision a person is making in the present) (Intentional Decisions Given Determinism)
- (2) No one can do anything that would be able to falsify a proposition expressing the completed state of the world in the remote past (e.g., a time in the Jurassic period) in conjunction with a proposition expressing a completed set of the laws of nature (Assumption).
- (3) Therefore, no one can do anything that would falsify a proposition (α) expressing a decision a person is making in the present. (1,2, Beta-Best)
- (4) Necessarily, if there is nothing that anyone can (ever) do that would falsify a proposition (α) expressing a decision a person is making in the present, then no one can have free will with regards to α (Assumption from ALTERNATIVE-POSSIBILITIES).
- (5) No one can have free will with regards to the decisions persons make in the present (α). (From 3,4, MP).
- (6) Necessarily, if proposition expressing the completed state of the world in the remote past (e.g., a time in the Jurassic period) in conjunction with a proposition expressing a completed set of the laws of nature necessarily entails the completed state of the world in the present (including some decision a person is making in the present, α), then no one can

have free will with regards to the decisions persons make in the present (α). (2-6, Conditional Proof).

And, we can put (6) slightly more informally as,

(6*) Necessarily, if determinism is true, no one can have free will.

This is the P-MODAL CONSEQUENCE ARGUMENT as I have articulated it.

Conclusion

This chapter has introduced the modal consequence argument generally, and Pruss' modal consequence argument (the P-MODAL CONSEQUENCE ARGUMENT) in particular. In 2.1, I outlined the informal consequence argument and suggested that the argument is powerful because the involved principles are so intuitive. In 2.2, I articulated the consequence argument into a modal and formal way and explained why McKay and Johnson's coin flipping example (1998) was such a powerful objection to the modal consequence argument. However, in 2.3, I showed that by incorporating Pruss' proof for Rule Beta and his modal definition of the N-Operator, we can show that McKay and Johnson's objection fails. Further, in 2.4, I explained a related objection from Pereboom and McKenna (2016) that threatened Pruss' particular formulation. I argued that there is likely no need to disjoint the fixity principles because the motivation is plausibly similar for both fixity principles.

The first three chapters of this thesis is meant to articulate the main argument that this thesis will defend, which is an argument for the incompatibility of physicalism and free will. In chapter one, I defined three key terms: (I) physicalism, (II) determinism, and (III) free will. In this chapter, I have explained how the modal consequence argument can be used to defend the incompatibility

determinism and free will. In the next chapter, I shall use the framework the consequence argument to articulate a new argument for the incompatibility of *physicalism* and free will.

CHAPTER THREE. THE B-MODAL CONSEQUENCE ARGUMENT

Introduction

In Chapter Two, I mainly focused on the validity of the P-MODAL CONSEQUENCE ARGUMENT. In this chapter, I am going to consider a new modal consequence argument. This argument, which I shall refer to as the B-MODAL CONSEQUENCE ARGUMENT, is an argument against the compatibility of free will and non-reductive physicalism (which is a subset of Type-B physicalism, hence the B in “B-MODAL CONSEQUENCE ARGUMENT”). The B-MODAL CONSEQUENCE ARGUMENT is structured in a way that relies on BETA-BEST just like the P-MODAL CONSEQUENCE ARGUMENT. The main purpose of Chapter Three is then to simply lay out and explain the B-MODAL CONSEQUENCE ARGUMENT. As I see the argument, there are two vulnerable or controversial premises in the argument. Subsequent chapters will provide defenses of these two premises before discussing the implications of the B-MODAL CONSEQUENCE ARGUMENT.

I am going to begin (§3.1) by discussing the worthwhileness of considering physicalism at all. This will (hopefully) help the reader to understand why I am interested in defending a new version of the consequence argument instead of the P-MODAL CONSEQUENCE ARGUMENT, which is an argument against the compatibility of determinism and free will. In §3.2, I will lay out my new version of the consequence argument (i.e., the B-MODAL CONSEQUENCE ARGUMENT). This chapter will set up the next part of the thesis where I will defend the potentially controversial elements of the B-MODAL CONSEQUENCE ARGUMENT.

3.1 The Relevance of Physicalism

I ended Chapter Two by defending the validity of the P-MODAL CONSEQUENCE ARGUMENT. This was the version of the consequence argument that concludes that determinism and free will are incompatible. In my discussion in Chapter Two, I did not give extensive defense the potentially controversial individual premises (other than noting that the two fixity principles were pre-theoretically intuitive). I focused, instead, on the controversial modal transfer rule (i.e., Rule Beta). For the sake of my point here, let us suppose, for the moment, that the P-MODAL CONSEQUENCE ARGUMENT is sound (i.e., all its premises are true) and convincing. That is, suppose we were convinced that free will was incompatible with determinism. Having assumed the argument is sound and convincing, can we learn anything about the nature of free will from this discovery? We would have learned that free will is related to the nature of lawhood in important ways: that certain kinds of physical laws (i.e., laws consistent with determinism) critically threaten the possibility of free will in any possible world. However, have we learned, for example, whether a robot could have free will? Or a whale? Or a human person (in the actual world)? What about a manipulated human person? Or, a human person with disabilities? The P-MODAL CONSEQUENCE ARGUMENT would certainly entail that a whale, a robot, and a human person (in the actual world) would all certainly *not* possess free will *in a deterministic universe*, but this fact is simply due to the nature of the laws governing the world are consistent with determinism and has no specific regards for the nature of whales, robots, or human persons in the actual world.

What made determinism relevant was the fact that the theory entailed there was a state of affairs (in conjunction with the laws of nature) that was (i) out of my (or any human person's) causal reach, and was (ii) logically entailing my (and every human person's) decision and action. It is (I) and (II) that made determinism relevant for the modal consequence argument. However, we can quickly see that there are other metaphysical theories that might plausibly entail (i) and (ii). Any

metaphysical theory that entails a state of affairs Σ is (I) out of my (or any human person's) causal reach, and (II) necessarily entailing my (and every human person's) decision and action will be subject to the modal transfer rule BETA-BEST, and a version of the modal consequence argument.⁴⁸

⁴⁸ Here is one such example: theism and, in particular, thomistic varieties of theism. David Oderberg (2016) argues that thomistic theists should adopt a doctrine he calls *Premotion*. Premotion is a doctrine roughly asserting “the impossibility of our doing anything without the assistance of God” (Oderberg, 2016, p. 208). Premotion states that God predetermines all human decisions in the sense that if God predetermines Jane to attend Birmingham (as opposed to Durham) for University, then necessarily Jane will attend Birmingham, *however* (and this is the key aspect) God's predetermining Jane's decision does not affect the modal status of Jane's decision (Oderberg, 2016, p. 209). As Oderberg notes, “If an effect [Jane's decision] is contingent, that it is moved by God makes no difference to its contingency. If an action is both contingent and free, as in the case of human or angelic free will, it remains free albeit predetermined infallibly to the act it performs” (ibid). If Jane contingently chooses to attend Birmingham, then, regardless of God's predestination, Jane freely and contingently chooses to attend Birmingham. So, God's premotioning is a variety of God's predestinating that preserves the modal status of the individual states or events that God predestines. Notice, though, if we grant the coherence and conceivability of such a doctrine (i.e., divine premotion), then we have a particular theory of theism (a metaphysical system) where a state of affairs Σ (e.g., God's premotioning or decreeing Jane to attend Birmingham) is (I) plausibly outside of Jane's causal reach, and (II) Σ logically or metaphysically necessitates Jane's decision. Thus, the modal consequence argument could be used to suggest free will is actually incompatible with such a theistic metaphysical system (i.e., one that involves this sort of premotion). Oderberg (2016, p. 217) foresees this possibility in the following passage:

Yet why, precisely, should this be a problem? Here, no doubt, the critic will invoke one of the various ‘transfer’ principles much discussed in the free will literature (on which, for an interesting recent contribution, see Pruss 2013). He might propose an operator such as ‘X never has a choice about whether or not P’, or ‘X has no power causally to falsify P’, for some agent X and proposition P. He might then argue as follows. If God moves S to perform A, for particular agent S and action A, then S never has a choice about whether or not God so moves him. (Reason: God's total independence of all creatures.) If God moves S to perform A, this logically entails that S performs A. (Reason: the infallible efficaciousness of the divine decree.) By transfer of necessity (the necessity of the operator just introduced), S never has a choice about whether or not to perform A. So it seems that divine predetermining premotion necessitates the consequent after all, contrary to the insistence of Thomists that this is not the case.

This is an interesting *new* version of the modal consequence argument. Whether or not Oderberg can successfully defend his variety of theism from this argument is not my concern here. But, for the sake of my broader point, assume it is successful for a moment. A modal consequence argument of the sort Oderberg describes would give us a sort of information about free will that the P-MODAL CONSEQUENCE ARGUMENT did not. It would tell us that whales, ghosts, and human persons (in the actual world) could not have free will *if there exists a God predestining all things (in the manner Oderberg describes)*. This is informative to an extent, however probably not as informative as one wishes. Other philosophers have since raised similar versions of modal consequence argument against other varieties of theism. See Roger Turner and Jordan Wessling (Wessling & Turner, 2021).

Physicalism does not obviously take a stand on whether determinism is true or false. Physicalism, like determinism, is a thesis about the metaphysical structure of the world. However, physicalism is also a thesis about the nature and metaphysical structure of human persons and, notably, phenomenal consciousness. If determinism is incompatible with free will, then whales, robots, and human persons cannot have free will if the laws of nature are deterministic. If physicalism is incompatible with free will, then whales, robots, and human persons cannot have free will because of the kind of beings they are: purely physical beings.

To use David Chalmers' metaphor, if physicalism was incompatible with free will, this would entail that, if God wanted to make free creatures, he would need to reach to the non-physicalist shelf of ingredients at some point in creation (Chalmers, 1996, p. 41). If this is the case, then we can ask: *What would plausibly be this non-physicalist ingredient?*⁴⁹

3.2 The B-Modal Consequence Argument

In this section I am going to outline the B-MODAL CONSEQUENCE ARGUMENT. The B-MODAL CONSEQUENCE ARGUMENT is a version of the modal consequence argument (broadly speaking) which concludes that non-reductive physicalism and free will are incompatible.

⁴⁹ One might wonder whether all of these modal consequence arguments stand or fall together. Certainly, if the BETA-BEST rule was shown to be invalid, they would fall together. However, they are all different in key respects. The potential modal consequence argument against physicalism and free will would have certain similarities to the P-MODAL CONSEQUENCE ARGUMENT. However, there would be key differences. Here is an example of a potential key difference: the possibility of time-travel. A key premise of the P-MODAL CONSEQUENCE ARGUMENT, which is usually thought to be intuitive, is that persons cannot change events in the remote past. However, the possibility of time-travel could conceivably threaten that premise. But, time-travel has no clear bearing on the relationship between physicalism and free will. An example like this would be relevant to one argument, but not the other. There are objections to the P-MODAL CONSEQUENCE ARGUMENT that will be relevant to the modal consequence argument against physicalism and free will (see §5.1-5.2).

I will begin this section reviewing the basic assumptions involved in the modal consequence argument (i.e., the assumptions that are common to both the B-MODAL CONSEQUENCE ARGUMENT and the P-MODAL CONSEQUENCE ARGUMENT). Subsequently, I will explain two novel principles that are unique to the B-MODAL CONSEQUENCE ARGUMENT. Finally, I will present the formal argument.

3.2.1 Principles Common To Both (The P and B) Modal Consequence Arguments

The most helpful way to begin is to review some of the common principles assumed in the B-MODAL CONSEQUENCE ARGUMENT, which are also assumed in the P-MODAL CONSEQUENCE ARGUMENT. To this end, I will first briefly review in this section BETA-BEST, ALTERNATIVE-POSSIBILITIES, FREE WILL, the N-Operator, and the F-Operator.

The argument I am going to present uses the logical structure from the P-MODAL CONSEQUENCE ARGUMENT, which was discussed in 2.3, and, thus, uses Pruss's modal principle known as BETA-BEST:

BETA-BEST: If there is nothing anyone can do that would falsify p , and necessarily p entails q , then there is nothing anyone can do that would falsify q .

For example, if there is nothing I can do to falsify "Clark Kent does not really exist," and this proposition necessarily entails "Superman does not really exist," then there is nothing I can do to falsify "Superman does not really exist."

Additionally, I am going to utilize Pruss' understanding of the N-Operator where Nr is true *iff* there is nothing that anyone can (ever) do that would falsify r (From Pruss, 2013, 433-434). For example, no one cannot falsify the proposition that Saturn has seven rings (D), and so we can infer

that ND is true. Following Chisholm and Plantinga, I assume that propositions and states of affairs are very closely linked (Plantinga, 1978; Chisholm, 1967). In particular, I assume that if you cannot falsify the proposition:

(1) Saturn has seven rings,

then you cannot make the state of affairs

(2) Saturn having seven ring

not obtain. So, we can also express the N-operator in the following way, Nr is true *iff* there is nothing that anyone can ever do that would make a state of affairs r not obtain. I assume these are equivalent statements. The N-operator is necessary for expressing the BETA-BEST principle more formally.

For simplicity, I am also going to use an additional operator (just like I did in Chapter Two) called the F-Operator, which indicates that I can *freely will* some particular outcome (e.g., I can freely decide to marry Emily).

In the F-Operator that I will use, I am using my definition of free will from Chapter One:

FREE WILL: A person P has free will *if and only if* (I) P has the ability to consciously choose an outcome ϕ (i.e., form an intention to ϕ or make a decision to ϕ) where ϕ is distinctly caused by P 's conscious or mental state or event, and (II) when P made a decision to ϕ , P could have chosen $\neg \phi$.

According to this conception of free will, I am, therefore, assuming the principle ALTERNATIVE-POSSIBILITIES.

ALTERNATIVE-POSSIBILITIES: A creature C is free in ϕ -ing *only if* C , when choosing to ϕ , could decide to $\neg \phi$.

In summary, BETA-BEST, ALTERNATIVE-POSSIBILITIES, the N-Operator, and the F-Operator are all elements of both varieties of the modal consequence argument. Next, I will overview the two new principles relevant for my argument.

3.2.2 Principles Only In The B-Modal Consequence Argument

There are two novel principles in the B-MODAL CONSEQUENCE ARGUMENT that are not present in the P-MODAL CONSEQUENCE ARGUMENT. The first new principle is simply a proposition or state expressing some present intentional decision in a possible world where non-reductive physicalism is true. In order to make this more concrete, imagine the following scenario:

Faculty Meeting: Emily is a new confident faculty member. At her first faculty meeting, her department head asks a genuine question from the faculty and Emily knows the answer before her colleagues respond. However, being new, Emily is afraid to answer. After seeing no one else raises their hand in a reasonable amount of time, Emily raises her hand to answer the question.

In this situation, we can use E as a proposition expressing Emily raising her hand at that moment (t). In order to imagine E in a world where non-reductive physicalism is true, it will be helpful to review non-reductive physicalism.

In §1.3.1 and 1.3.3, I defined non-reductive physicalism as consisting of (i) GLOBAL SUPERVENIENCE, (ii) NOMOLOGICAL CLOSURE, and (iii) MENTAL ONTOLOGICAL AND CAUSAL

IRREDUCIBILITY. The combination of (i)-(iii) is what I am referencing for non-reductive physicalism.

To make sense of **Faculty Meeting**, recall that, in the dot-mosaic analogy (from §1.3.1), one uses tiny dots to paint a more complex picture or pattern (c.f., Lewis, 1986, p. 14). In this scenario, one can think of the dots as analogous to fundamental microphysical configurations (i.e., fundamental microphysical objects and their features; e.g., electrons or field-states) and the mosaic pattern as analogous to the macrophysical configuration in question. Using this dot-mosaic analogy, the completed macrophysical state of the world at t would be the grand mosaic in its entirety, and we can refer to this completed state as C_t . We can also refer to the completed state of fundamental microphysical world (referring to the same world as the prior sentence) at t as F_t , which would be analogous to entire configuration of dots on the mosaic. To continue the analogy, suppose we do not care about the entire mosaic but a portion of it; in that case, it is still true that the painting and all its features (e.g., properties, states, events) *in that one portion* of the mosaic (e.g., a bird in the corner) depends on the specific dots *in that region* (e.g., the dots specifically making up the bird).⁵⁰

For the proponent of MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY the relevant specific macrophysical state in **Faculty Meeting** is Emily's deciding to raise her hand at t (E). While it is true that the completed microphysical state at t , F_t , entails the completed macrophysical state at t ,

⁵⁰ One small clarification, one might think about macrophysical properties that deal with more than just this region, for example if I am looking at a particular bird in the corner of the mosaic, the bird could have the property of being 10 inches from the tree (in the other corner of the mosaic). *That* property would depend on more than just the particular dots undergirding the bird directly. This is analogous to certain macrophysical properties in the world, but will not be relevant to my analogizing point here.

C_t , which entails E, it is additionally true that there is a *specific* microphysical configuration realizing E (i.e., Emily willing to raise her hand at t). Let us simply refer to the specific (token) microphysical configuration that specifically realizes E as S. Back in §1.3.3, I explained that the proponent of MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY can understand realization in terms of grounding or constitution (or another non-identity relation), but the key conditions for realization (Baysan, 2019, 79) are that (I) S metaphysically necessitates E, (II) S is more fundamental than E, and (III) S and E are not identical.

So, S is the specific microphysical configuration that entails Emily's deciding to raise her hand at t . S is analogous with the specific dots undergirding a specific area of a dot-mosaic. In the dot-mosaic example, there is no one specific configuration that the dots have to be in in order for the macrophysical image to be present. If one dot was removed or if another was slightly moved over, the bird-image would presumably still be present in the mosaic. In that sense, the bird image can be necessarily entailed by the token configuration of dots, but there can be multiple dot-configurations that would be sufficient for the bird-image. This is because the bird image is what philosophers call *multiply realizable*; multiple configurations could be sufficient for the bird image, but there are parameters on how much one can mess with the dots without changing the image.⁵¹ However, if the image is changed at all the dots must have been changed. It is the same for Emily's raising her hand. E is multiply realizable by different microphysical configurations,

⁵¹ Presumably, you can completely replace the dots with some other medium or quantity and the bird would still be present. For example, if you put micro-images of birds in place of dots (assuming the images were small enough), you would still have the macro-image of the bird. I believe Hilary Putnam was the first to point this multiple realization argument in his (1967), however multiple philosophers have argued this since.

but (I) there is a range of configurations that will realize E and (many, many more) configurations that will not realize E and (II) there is a token-specific configuration that actually does realize E in the world.

With this information we can thus state our first principle:

INTENTIONAL DECISIONS GIVEN NON-REDUCTIVE PHYSICALISM: Necessarily, assuming non-reductive physicalism is true, a state of affairs E expressing a particular intentional decision at t (e.g., Emily deciding to raising her hand at t) is metaphysically necessarily entailed by a state of affairs S expressing a configuration of microphysical objects and properties (i.e., $\Box (S \rightarrow E)$).

By metaphysical necessity (e.g., grounding, constitution, realization), Emily's deciding to raise her hand at t (E) will have a token microphysical configuration at t (S) that necessitating it. This is the first novel principle for the main argument of this thesis and is very similar to the manner in which Van Inwagen defined determinism in his version of the consequence argument (see §2.1-2.2).

The second novel principle deals specifically with Emily's causal relation to S. And, here, simply intuitively, it would be odd to suggest that Emily would have a choice with regards to this microphysical configuration undergirding her decision (S). She does not have a choice over which specific configuration is true in the actual world (as opposed to another sufficient configuration). She also does not have a choice on the parameters that constitute the S-type that is sufficient for E, or, as Cover and O'Leary-Hawthorne put it, "people don't have a choice about the exact spatial relations among electrons and other microphysical particles making up their microstates" (Cover

& O'Leary-Hawthorne, 1996, p. 59).⁵² Or as Pearce similarly remarks: we do “not appear to have control over, for instance, the values of the electron field in the region of [our] brain[s]” (Pearce, 2022, p. 1109). It is not like Emily, at *t*, was thinking to herself, ‘maybe I should raise my hand *or* I could change the particular configuration of microphysical particles (or the state of a quantum field)!’ No, of course not, it would be peculiar to ever have such a thought. If we take that intuition seriously, we can suggest that the following is true

FIXITY OF THE MICROPHYSICAL CONFIGURATION: Necessarily, assuming non-reductive physicalism, there is nothing anyone can do to make *S* not obtain (FIXITY OF THE MICROPHYSICAL CONFIGURATION)

This principle is supposed to be analogous to the Fixity of the Past or Laws in the original modal consequence argument.

Before moving forward, one helpful clarification might be needed. One might think of the situation Emily is in (with relation to *S*) as the relation one has with the temperature in a room. For example, in the room I am currently in, I do not appear to have a choice over the state of kinetic energy in the room. However, I believe I can change the thermostat, which would change the state of kinetic energy.⁵³ Does this analogy inspire a kind of objection or problem to the FIXITY OF THE

⁵² Cover and O'Leary-Hawthorne (1996) are, to my knowledge, the first to attempt applying the consequence argument to materialism or physicalism.

⁵³ Bailey and Rasmussen (2020) bring up a similar analogy that they consider: “We might wonder where one could be responsible for unknown things that are *involved* in ordinary activity. For example, maybe you want to bake some

MICROPHYSICAL CONFIGURATION? The fact that I can change the thermostat might be a problem for FIXITY OF THE MICROPHYSICAL CONFIGURATION if the reasoning for this principle is purely inspired from intuition. However, I think there is a good argument, apart from intuition, for FIXITY OF THE MICROPHYSICAL CONFIGURATION which I will discuss in Chapter Four.

With these two principles, I believe we can form a new version of the modal consequence argument. I will explain this argument next.

3.2.3 The Formal Argument

The previous two sections, I have sought to show the reader the two new principles: INTENTIONAL DECISIONS GIVEN NON-REDUCTIVE PHYSICALISM and FIXITY OF THE MICROPHYSICAL CONFIGURATION. Taking S to be the particular microphysical configuration at t that metaphysically necessitates Emily's deciding to raise her hand at t , E , (because S grounds, constitutes, or realizes E), we can formulate the B-MODAL CONSEQUENCE ARGUMENT:

- | | |
|--------------------------------|------------------|
| (1) $\Box ((S) \rightarrow E)$ | Assumption |
| (2) NS | Fixity Principle |
| (3) NE | 1,2, Rule Beta |
| (4) $NE \rightarrow \neg FE$ | Assumption |
-

break, yet you don't know how exactly the dough rises or how the atoms in your fingers will move as you knead. In this case, perhaps you could be responsible for the resulting break, even when the micro-physical processes are unknown. That is plausible. We find it significantly less plausible, however, that you are responsible for those unknown microphysical activities involved in baking. You are, rather, off the hook when it comes to the processes themselves precisely because you are innocently ignorant of them" (205). I think I am in agreement with the, although they are making a slightly different claim about ignorance and moral responsibility.

(5) $\neg$ FE

(4,3, MP)

Here, N stands for the N-Operator (e.g., NS stands for there is nothing anyone can do to make S not obtain). In ordinary discourse, the argument can be stated as follows (where I will use Emily's hand-raising as the paradigmatic example again):

- (1) Necessarily, assuming non-reductive physicalism is true, a state of affairs E expressing Emily deciding to raising her hand at t is metaphysically necessitated by a state of affairs S expressing a particular microphysical state (i.e., $\Box (S \rightarrow E)$). (INTENTIONAL DECISIONS GIVEN NON-REDUCTIVE PHYSICALISM)
- (2) Necessarily, assuming non-reductive physicalism, there is nothing Emily (or anyone) can do that would make S not obtain (FIXITY OF THE MICROPHYSICAL CONFIGURATION)
- (3) Therefore, there is nothing that Emily (or anyone) can do that would make Emily's deciding to raise her hand at t not obtain (From 1,2, BETA-BEST)
- (4) Necessarily, if there is nothing that Emily (or anyone) can do that would make Emily's deciding to raise her hand at t not obtain, then Emily cannot freely decide to raise her hand at t . (Assumption from ALTERNATIVE-POSSIBILITIES).
- (5) Emily cannot freely decide to raise her hand at t . (From 3,4, MP).
- (6) Necessarily, if Emily's deciding to raise her hand at t is necessarily entailed by a proposition S (expressing a particular fundamental microphysical state at t), then Emily cannot freely decide to raise her hand at t . (2-6, Conditional Proof).

Looking closer at this argument, the most controversial premises are premise 2 and premise 4, but I will spend a moment explaining the other premises to show why they are not plausible candidates for potential objections. Premise 1 is merely an entailment from non-reductive physicalism to

Emily's situation. The third premise of this argument is, in effect, a fixity principle, but it follows from (1) and (2). To help distinguish (3) from (2), I will refer to (3) as **FIXITY OF THE PRESENT**:

FIXITY OF THE PRESENT: Therefore, there is nothing that Emily (or anyone) can do that would make Emily's deciding to raise her hand at t not obtain (From 1,2, BETA-BEST)

Of course, this fixity principle is counterintuitive taken by itself. **FIXITY OF THE PRESENT** entails that Emily has no other (exercisable) ability at t other than the ability to raise her hand. Thus, **FIXITY OF THE PRESENT** entails Emily never had an ability to perform any action which she did not actually perform, which is a highly counterintuitive claim (I will discuss this further in §5.1).

If Emily cannot perform any action other than she actually performs, then becomes similar to Alvin Plantinga's Mr. McEar (Plantinga, 1967, p. 170). According to Plantinga, Mr. McEar is a hypothetical person with a very interesting set of abilities. Normally we think that people have a large set of different abilities (e.g., I have the ability to be a doctor one day or be a fighter-pilot, etc), but McEar essentially only has one ability: the ability to scratch his right ear. In every possible world where McEar exists, he only scratches his right ear (all the time?). McEar cannot do anything that would make his itching his ear not obtain (i.e., he cannot stop scratching his right ear). In that sense, McEar and Emily are analogous. However, there is at least one interesting difference. In the Emily example, Emily's set of exercisable abilities always has one constituent: like McEar's set of exercisable abilities. But, Emily's particular single constituent does change moment by moment: unlike McEar. At t , Emily can raise her hand, but the moment after t , Emily can speak, and so on. Intuitively, McEar does not have free will. So, **FIXITY OF THE PRESENT** is radical in that, if it were really true of all abilities in the actual world, it would conflict with a lot of our ordinary beliefs about what we can do any given moment. We normally think we can exercise multiple abilities in

a given choice-situation. Assuming ALTERNATIVE-POSSIBILITIES, premises 5 and 6, then, simply logically follow from the previous premises.

The last significant step of my argument is the following premise (6), which we can re-state for simplicity and give a proper name:

INCOMPATIBILIST CLAIM ($\neg$ FC): Therefore, if non-reductive physicalism is true, no person can have free will.

Here, I am using F as a symbolic designation. The ultimate conclusion of the conditional proof can be stated more simply that, if non-reductive physicalism is true in some possible world w , then not even God could create a free creature in w , and so non-reductive physicalism and free will are incompatible.

Conclusion

This chapter explained the B-MODAL CONSEQUENCE ARGUMENT, which will be focus of the rest of the thesis. I first explained the common principles between the P-MODAL CONSEQUENCE ARGUMENT and the B-MODAL CONSEQUENCE ARGUMENT. Then, I briefly explained the two novel principles in the B-MODAL CONSEQUENCE ARGUMENT before laying out both the formal and informal versions of the argument. As I have presented the argument, there are really two controversial aspects. The first controversial aspect is the FIXITY OF THE MICROPHYSICAL CONFIGURATION. This is where I will turn next in the thesis.

CHAPTER FOUR. THE FIXITY OF THE MICROPHYSICAL CONFIGURATION

Introduction

In this chapter, I will discuss the second premise of the B-MODAL CONSEQUENCE ARGUMENT. My main aim is to defend the crucial fixity principle. The second premise of B MODAL CONSEQUENCE ARGUMENT is what I call the FIXITY OF THE MICROPHYSICAL CONFIGURATION principle, which revolves around a very specific fundamental microphysical state.⁵⁴

The plan for this chapter is as follows: First (in 4.1), I will present my main argument from NOMOLOGICAL CLOSURE. Then, I will discuss two potential responses to this argument. The first is from Christian List (2019a; 2019b), and the second is from Jessica Wilson (2021). I will demonstrate that List's potential response is actually consistent with my argument. Finally, in §4.4, I will argue that Wilson's subset view of mental causation is not plausible, and, therefore, not convincing as a potential counterargument.

4.1 The Fixity of the Microphysical Configuration

This section will explain the main argument for the FIXITY OF THE MICROPHYSICAL CONFIGURATION principle. I am going to begin with an analogy involving the magician Merlin to

⁵⁴ Not as catchy as "Fixity of the Past" but the point comes across.

set up the main argument. My argument is inspired from L.A. Paul's problem of additivity (2007, p. 275).

To begin, consider the following (analogous) scenario

Merlin Scenario: Imagine that the wizard Merlin jams a sword (i.e., Excalibur) in a stone next to my house, and, subsequently, Merlin casts a spell making it the case that only King Arthur can pull the sword out of the stone. Further, imagine that Arthur will come into existence long after I die. Perhaps I lived from the year 1993 to the year 2093, and Arthur was born on Earth in the year 2500.

Of course, I can normally move (ordinary) objects in my yard. If a ball or some construction materials flew into my yard, then I would naturally believe I could move the object(s) and take action to get it out of my yard. So, in the Merlin Scenario, I would naturally believe I could quickly move Excalibur (the sword) like any other object in my yard. Here is the question I want to put forward: what can I do that would make Excalibur's being stuck in a stone in my yard not obtain? Of course, if Excalibur were an ordinary sword, under normal circumstances, I could simply use my hands to pull it out of the stone (or rent a bulldozer), which would make the relevant state not obtain. But, here, what could I do?

The problem that is easily discernable is that Merlin's magic makes it the case that *only Arthur* can pull the sword out of the stone. So, if I was Arthur, *then*, I would be able to pull Excalibur out of the stone in my yard. Arthur has an easily imaginable ability he could potentially exercise here.

However, I am not Arthur. Given, the magic in the background of the Merlin Scenario, therefore, it is impossible for me to do anything that would make the state of Excalibur being stuck in a stone in my yard not obtain. I have no such ability.

Now, consider the scenario I brought up in Chapter Three:

Faculty Meeting: Emily is a new confident faculty member. At her first faculty meeting, her department head asks a genuine question to the faculty, and Emily knows the answer before her colleagues respond. However, being new, Emily is afraid to answer. Finally, after seeing no one else raises their hand in a reasonable amount of time, Emily decides to raise her hand to answer the question and promptly raises her hand.

In this scenario, assume that Emily can, like ordinary people, exercise all sorts of abilities under normal circumstances.

If we assume non-reductive physicalism is true in Faculty Meeting, we can apply infer the following is true

- (1) Necessarily, assuming non-reductive physicalism is true, a state of affairs E expressing Emily deciding to raising her hand at t is metaphysically necessitated by a state of affairs S expressing a particular microphysical state (i.e., $\Box (S \rightarrow E)$). (INTENTIONAL DECISIONS GIVEN NON-REDUCTIVE PHYSICALISM)

Now, assume the following physicalist doctrine true:

NOMOLOGICAL CLOSURE: for any fundamental microphysical state, the objective probability of that state obtaining is identical to the probability solely given prior microphysical states of the universe in conjunction with the fundamental laws of nature.

As I discussed in Chapter Two, objective probability means the likelihood of the event or state given *all* antecedent factors and the laws of nature. For example, suppose ghosts exist in Emily's Faculty Meeting scenario and are not physical entities. If this is the case, S's objective probability

must account for the possibility that a ghost would do something that would make a difference toward whether or not S obtains.⁵⁵ If we then apply Nomological Closure to Emily, we can infer the following:

- (2) Necessarily, the fundamental microphysical state S is such that its objective probability is identical to its probability solely given prior microphysical states of the universe in conjunction with the fundamental laws of nature.

To continue using the ghost example, there are three relevant probability statements: First, the probability of S solely given prior microphysical states and the laws of nature (which I will symbolize as $(P \wedge L)$), second, the objective probability of S is the probability of S given all factors (which I will symbolize as Σ), and, third, there is the probability of S given both $(P \wedge L)$ and the Ghost's possible interference (Φ). NOMOLOGICAL CLOSURE suggests that all three probabilities are identical, as can be seen below:

$$\Pr\left(\frac{S}{P \wedge L}\right) = \Pr\left(\frac{S}{\Sigma}\right) = \Pr\left(\frac{S}{(P \wedge L) \wedge (\Phi)}\right)$$

The probability does not have to be 1 (e.g., it could be .2432 or .78), but each calculation's probability must be identical to each other for NOMOLOGICAL CLOSURE to be true. NOMOLOGICAL CLOSURE constrains the kind of potential difference-makers to S as either (I) fundamental microphysical difference-makers or (II) non-fundamental microphysical difference-makers that

⁵⁵ However, if ghosts exist *and* NOMOLOGICAL CLOSURE is true, then either ghosts have no causal ability that could make a difference toward S *or* their ability to make a difference toward S is somehow accounted for in the probability of S solely given the prior microphysical states of the universe in conjunction with the fundamental laws of nature.

are nomologically invisible (i.e., do not *add* to the objective probability of S beyond the probability of S given S's prior fundamental microphysical states and the fundamental laws of nature). In the case of the ghost, we would likely infer that ghosts are impossible or epiphenomenal.

In the case of Emily, there is one more important non-reductive physicalist doctrine to consider:

MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY: Some mental entities and features are (I) *not identical* to the configurations of fundamental microphysical objects and features they depend on and are (II) distinctively causally efficacious with respect to the configurations of fundamental microphysical objects and features they depend on (Paraphrased from Wilson, 2021, pp. 1-2).

This is relevant because all of Emily's potential intentions or decisions are mental states (as opposed to physical states). Additionally, based on the view of action presented in §1.1 (i.e., actions are events relevantly caused by mental events), all of Emily's actions necessarily entail distinct mental causes. Thus, we can say in this scenario that:

- (3) Anything Emily could possibly do would be a mental event or necessarily involve a mental event, which is essentially not identical to a fundamental microphysical event.

From (1) - (3), we are left realizing that Emily's abilities are inherently mental and S is nomologically closed (and additionally that S necessitates Emily's decision at *t*).

So, what kind of ability could Emily have exercised that would make S not obtain? This is the same kind of question we asked in the Merlin Scenario, when we asked what kind of ability I could have exercised that would made the sword being stuck in the stone not obtain. We assume from

our pre-theoretic intuition that Emily *can exercise all sorts of abilities*. Emily can jump, sing, or ride a horse into that faculty meeting.

However, in this Faculty Meeting Scenario, it seems clear (via 3) that Emily's actions and decisions are not fundamental microphysical events. Thus, if Nomological Closure is true (which is necessary given non-reductive physicalism), then we can suppose,

- (I) If Emily has some ability (ϕ) that could make S not obtain, then Emily's exercising of ϕ would be causally overdetermining $\neg S$ in a fashion that is nomologically invisible.

In other words, Emily's ability to ϕ can not *add* to the objective probability of S beyond the probability of S given S's prior fundamental microphysical states and the fundamental laws of nature). However, before moving forward I need to explain what I mean by overdeterminism here.

Following L.A. Paul (2007, pg., 269-270), I understand overdeterminism as inherently a causal phenomenon. We can roughly characterize a phenomena as being overdetermined if two separate sufficient causes produce the same effect. To take an oft-used example, if Jack and Sally both throw a rock at a window and the window breaks, the window breaking is potentially a case of overdeterminism.⁵⁶ It seems like both rocks or rock-throwings cause the window to break. In order for it to be a genuine case of overdeterminism (rather than just potentially a case), neither rock could hit the window before the other; they both must strike at the precise moment the other strikes (otherwise, it would be preemption). Additionally, the window must not have been so strong that

⁵⁶ I believe Paul (2010) was the first place I saw this type of example. Bennett (2003) uses a firing squad example to get the same point across.

it would have required two rocks (as opposed to one rock) thrown at it in order for it to break (otherwise, the two rock-throwings would be a joint-single-cause). Overdeterminism rather is a case “involving multiple distinct causes of an effect where the causation is neither joint, additive, nor preemptive (and it is assumed the overdetermining causes do not cause each other)” (Paul, 2007, pp. 269-270). Using the Jack and Sally example, each rock “makes exactly the same causal contribution as the other causes to the effect (so the causal overdetermination is symmetric); each cause without the others is sufficient for the effect; and for each cause, the causal process from cause to effect is not interrupted” (ibid.). The key for such a case is that each cause is *distinct* and *symmetric*, but, importantly, not *jointly causal* or *additive* (e.g., the window cannot be strong enough to withstand one rock but not two rocks).

Because there are several conditions for a phenomena to be genuinely overdetermined, it seems like the amount of genuinely overdetermined phenomena in the world should be rare. Take Jack and Sally’s rock-throwings, it is theoretically possible that the window breaking was a genuine case of overdeterminism. However, in any actual case like where Jack and Sally throw rocks at a window, the odds of the window shattering being genuine overdetermined (as opposed to preempted) is very low. For if either rock hits the window before the other, even a micro-second before the other, it would not be overdeterminism.⁵⁷

If Emily has some ability (ϕ) that could make S not obtain, then there would be quite a lot of overdeterminism in the world. This seems implausible to some philosophers (although, not all).

⁵⁷ This is *prima facie* a problem for anyone willing to accept that genuine overdeterminism happens commonly.

As L.A. Paul notes, the notion that there is widespread overdetermination “seems to produce a lot more causation in the world than we ever knew about or wanted” (Paul, 2007, p. 272). If every microphysical event or state that realizes a decision is overdetermined, there is more causation than we intuitively knew. It would, in essence, posit overdetermination as a very common occurrence, which seems problematic.

However, another reason this we should be skeptical that there is such overdetermination is that, if this is the case, then there are simply a lot of unexplained correlations between many causes. As Ted Sider notes (2003, 722),

Imagine a paranoiac who thinks that every time someone is shot, there are two causally independent shooters. He is crazy, but why? One reason (not the only one) is that it would be a coincidence that all these sharpshooters just happen to fire at the same places at the same times. This great regularity would need an explanation, and none could be given. [This] would require a massive, unexplained correlation between multiple causes.

This is certainly a reason to be skeptical of any claim of widespread overdetermination.

However, there is second reason it is implausible for Emily to have such an ability as ϕ (which is an ability such that, if she exercised it, would make S not obtain and is nomologically invisible to S's objective probability). The second reason is that the kind of overdetermination ϕ would imply seems notably different than overdetermination in the example of the Jack and Sally rock throwings. Jack and Sally's rock throwings overdeterminately breaking the window is unlikely to occur in front of our eyes, but theoretically it is possible because both rock throwings can be sufficient for the window to break while being distinct symmetrical events. It will be true that if

Sally had not thrown the rock, the window still would have shattered, and if Jack had not thrown it, the window would still have shattered. This does not seem true in Emily's case.

To see the difference clearly, consider this modified example from L.A. Paul (2007, p. 269):

Federer Case: Roger Federer, the tennis player, hits a tennis ball with his racket at t , transferring momentum (p) to the ball, changing its direction and flinging it across the net. Unbeknownst to Federer (but seeable to the crowd), a ghost appears right before Federer strikes the ball and strike the ball with an ethereal tennis racket at t . However, the tennis ball exhibits the same amount of momentum as it would have if only Federer had struck it.

We can ordinarily say that Federer *causes* the ball to fly through the air (via his intention to swing his racket).⁵⁸ Now, let us assume that the tennis ball flying off Federer's racket is overdetermined by Federer *and* the ghost. It is sensible to suggest that both the ghost's strike and Federer's strike are distinct sufficient events, and not jointly causing the ball to fly over the net. As long as the ghost times his strike right (or luckily), this case would not be a preemption case. All of this is conceivable, except the asymmetry of additivity. As L.A. Paul argues, "[t]his brings a central problem to the fore—the problem of additivity of conserved quantities. For example, under the laws, momentum is additive; when there are two transfers of momentum p the total transfer is $2p$. But the tennis ball only exhibits an increase of magnitude p , not $2p$!" (2007, p. 275). The key questions this analogy brings up is: (1) On what basis should we suggest we believe the Ghost

⁵⁸ Some theories of causation would suggest that the transference of a conserved quantity just is causation, but other theories do not make this strict of a claim. In this example, it does not matter what theory of causation one prefers.

really caused the ball to fly through the air at all, and (2) If Federer had not struck the ball at all (perhaps he swung and missed), and only the Ghost struck it, would the ball have sailed over the net? Taking the second question first, if the answer to (2) is affirmative, it is clear that the Ghost adds conserved quantities to the ball (i.e., momentum, change of direction). This contradicts the assumptions in the Federer Case, though, where Federer and the Ghost struck the ball together, and the Ghost's strike was not additive. Thus, the answer to (2) must be negative. But, now that we have established that the answer to (2) is negative, let us consider (1): Why think that the Ghost's and Federer's strikes are causally overdetermining in this case? In the Jack and Sally case, the rock throwings were, for example, distinct and symmetrical (i.e., each rock has enough momentum on its own to break the window). In the Federer Case, it is hard to see why one would naturally infer that the Ghost really caused anything. The more natural inference is that the Ghost only *appears* to be striking the ball, but is really epiphenomenal.

Emily is in much the same position as Ghost (and not like Sally) in relation to $\neg S$. Sally's rock-throwing seems causally efficacious regardless of whether or not Jack was present. Sally's presence must be accounted for in the objective probability of the window breaking at that moment. The Ghost cannot be accounted for in such a way. In Faculty Meeting, Emily's ability to refrain from raising her hand (which would make S not obtain if exercised at t) cannot be accounted for in the objective probability of S (like the Ghost). This second problem is worse for Emily than the first (where the first problem was that the addition of more causation to the world might make causation a vague concept). Overdeterminism seems imaginable in cases like Jack and Sally, but is very difficult to consider as plausible explanations in cases like Federer. And, Emily's case would be far more similar to the Ghost's relation to Federer than Sally's rock throwing to Jack's rock throwing. Thus, it seems like the following is what we should infer that:

- (II) It is not the case that Emily's exercising of ϕ would be causally overdetermining $\neg S$ in a fashion that is nomologically invisible.

If Emily's ability to ϕ does not *add* to the objective probability of S beyond the probability of S given S 's prior fundamental microphysical states and the fundamental laws of nature, then it

If we accept (I) and (II), then the obvious conclusion (via Modus Tollens) is

- (III) Therefore, it is false that Emily has some ability ϕ (where ϕ is an ability, if exercised, that would make S not obtain).

This (III) is the conclusion of my argument that uses Faculty Meeting. Cases like the Merlin Scenario and Faculty Meeting should lead us to question the plausibility of certain abilities, such as ϕ (within a non-reductive physicalist world).

This is my main argument for NS (The Fixity of the Microphysical Configuration). It is an argument that crucially assumes (1) NOMOLOGICAL CLOSURE and (2) MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY, which are necessarily entailments of non-reductive physicalism.

4.2 Christian List's Intra-Level Mental Causation

In the previous section, I argued for the NS claim: that S is true, and there is nothing Emily (or anyone) can do that would falsify S . In this section, I am going to explore a potential response to my argument from Christian List. List, like many non-reductive physicalists have previously encountered a similar argument in Jaegwon Kim's paradigmatic Exclusion Argument (1998; 2005). Consequently, it could be argued that List's reply to Kim has the potential to effectively counter my argument. Here, I am going to show that this is not the case. In fact, I will argue that my argument is perfectly compatible with List's response to Kim's exclusion argument and general

metaphysics of mental causation. To explain this, I am going to briefly explain the relevant part of Kim's argument and then go into List's response.

Jaegwon Kim argues for the following principle:

Bootstrapping Causation: A mental state (M) causes another mental state (M*) only if M causes M*'s microphysical realizer. (Kim, 2005, p. 40)

This principle is supposed to recognize the fact that in order for one mental state M at a given time t to causally impact a mental state M* at a subsequent time t^* , M would have to causally impact the microphysical state on which M* is dependent. To put a concrete example to this, imagine I have a desire (M) to eat some cheese at t , and then I make a plan (M*) a few moments later t^* to go to the refrigerator to get some cheese. Both my desire to get cheese (M) and my plan to go to the refrigerator (M*) are mental states that are metaphysically dependent on microphysical states (assuming GLOBAL SUPERVENIENCE). Kim's principle is simply suggesting that my desire to eat cheese must impact the microphysical state that my planning to go to the refrigerator depends on in some manner if my desiring cheese actually causally impacts my planning to get cheese.

If we accept Bootstrapping Causation, then there is a clear worry about NS (the Fixity of the Microphysical Configuration). Kim argues, as I did, that microphysics is closed. However, his argument goes further where his argument's goal is to show that "under the physicalist assumptions we are working with, mind-to-mind causation is in trouble just as much as mind-to-body causation" (2005, pg., 40). He continues (ibid):

Often the problem of mental causation is presented as that of explaining how the mental can inject causal influences into the causally closed physical domain, that is, the problem of explaining mental-to-physical causation. I wanted to do something more, namely to

show that physicalism can put in peril all forms of mental causation including mental-to-mental causation.

Kim's argument is presented in a few different fashions (e.g., 1998; 2005), but the relevant part of his argument for my explanation of List is his Bootstrapping Causation principle.

Christian List (2019) rejects the Bootstrapping Causation principle, and, by doing this, is able to reconcile mental causation with non-reductive physicalism. List's strategy can be explained by specifying two general theses:

- (1) Non-Reductive Layered World: The objects and features in the world are metaphysically and conceptually distinct from each other by being at different *layers*. Fundamental microphysical entities (e.g., particles and fields) are at the bottom, cells are at a different layer, chairs at an even different layer, and persons (and consciousness) at an even different layer. These layers are different because of the laws governing them and their level of description. (List, 2019a; 2019b; see also Oppenheim and Putnam 1958).
- (2) Difference-making counterfactual causation: x causes y iff (I) x is counterfactually sufficient for y and $\neg x$ is counterfactually sufficient for $\neg y$ and (II) x and y are at the same metaphysical level (e.g., a cell's process of mitosis can cause a cell duplicate to exist because they are both at the same level). (List 2019a; 2019b; List and Menzies, 2009).

The Non-Reductive Layered World (1) is just an extension of MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY to all distinct special science and non-fundamental objects and features.

Of course, all of these non-fundamental objects and features (e.g., chairs) are realized by fundamental microphysical configurations of objects and features.

In order to explain Difference-making counterfactual causation, List (List, 2019, pp. 133-134) borrows an example from Jackson and Petit (1990) to help illustrate his strategy:

Flask Heating: Imagine that you heat up a flask of water. When the water boils, the flask breaks due to the pressure.

List (2019, p. 133) asks, “What is the cause of the flask’s breaking?” There are two potential answers to this question. The first is that “the boiling of the water, which increases the pressure beyond the level the flask can withstand,” causes the flask to break (List, 2019, 134). The second is that “the underlying microstate of the millions of water molecules” causes the flask to break (ibid). These are two causal options for the flask’s breaking. (This resembles Emily in Faculty Meeting, and her choice to raise her hand quite nicely.) List argues that if we take a counterfactual or “difference-making” analysis of causation, we can apply a standard counterfactual test to both potential explanations. The standard counterfactual test is to ask if the following two counterfactuals hold for both causes:

1. Positive Counterfactual: If *C* were to occur, then *E* would occur
2. Negative Counterfactual: If *C* were not to occur, then *E* would not occur.

List (2019, p. 134) makes clear that if we take *C* to be the boiling of the water and *E* to be the flask breaking, then both the Positive and Negative Counterfactual seem true:

1*: If the water boils, then the flask would break.

2*. If the water does not boil, then the flask would not break.

Both 1* and 2* seem true because in the closest possible worlds where the antecedents occur, the consequents seem to also occur. However, we can also take *C* to be the underlying microstate of water (which realizes the water boiling), in the following manner:

1**: If the given microstate occurs, then the flask would break.

2**: If the given microstate does not occur, then the flask would not break.

List argues that 1** is plausibly true, but 2** is not. He explains:

The reason is that the boiling of the water is multiply realizable: many different microstates can realize the boiling. And many such microstates would equally make the flask break. So, the negative conditional for difference making would come out false. In particular, in some of the nearest possible worlds in which the given microstate does not occur, the breaking of the flask would still occur. (List, 2019a, 134).

He further notes,

An understanding of causation as difference making is central to the special sciences. Much of the methodology of inferential statistics, for example, revolves around the goal of determining which variables make a significant difference to which other, while holding fixed certain control variables...Using such methods, we have been able to find out that regularly taking a low dose of aspirin reduces the heart attack risk in certain kinds of patients, while eating lots of butter does not. We have learnt that air pollution makes a difference to cancer rates in the population, while the use of microwave ovens (as far as we know) does not... In each of these cases, scientists identify causation by looking for evidence of difference making, especially difference making that persists even when we

control other factors. Causes are difference makers, while anything that is not a difference maker is not a cause. It is hard to imagine what the special sciences would look like if we didn't have such reasoning at our disposal. (List, 2019, 131)

The notable fact about List's understanding of causation is that causation will only occur intra-level, not inter-level in metaphysical practice. If an effect E is at a certain metaphysical level L, then the cause of that effect will be within L. As Sydney Shoemaker, who holds a similar view, points out, the "advantage of this approach is that it provides a basis for saying that in some cases it is the instantiation of a mental property, rather than the instantiation of one of its realizer properties, that caused a certain effect, or contributed to causing it" (Shoemaker, 2003, 3). As L.A. Paul summarizes

Instead of true overdeterminism, there are multiple layers of causal relations, where the causal relations at higher layers non-reductively supervene on causal relations at lower layers. In this picture, there is no overdetermination because causal processes initiated by higher layer objects only cause higher layer effects and lower layer objects only cause lower layer effects... persons cause things like bouncings of tennis balls at their layer, while sums of cells cause things at theirs (Paul, 2007,p. 272).

In sum, according to List's view, causation is difference-making. Because of this, in conjunction with his layered worldview, causation only occurs intra-level and *never* inter-level.

On this view of causation and the layered word, it is quite clear why Kim's Bootstrapping Causation principle would not be true. To show this, reconsider my desiring cheese (a mental state M) at a certain time that supposedly causes (at least partially) my plan to go get some cheese out of the refrigerator (another mental state M*). According to Kim, in order for M to cause M*, it

would have to additionally cause (at least partially) M*'s microphysical realizer (e.g., P*). However, this is not true given List's two theses. The difference maker for P* would be a state or event on the same metaphysical level of the world (i.e., the microphysical level), just like the difference maker for M* would be a state at the psychological level of the world. According to this view, Kim's worry that mental causation is totally impossible given a physicalist closure doctrine does not hold true.

Given List's success against Kim, one might be tempted to apply this view to my argument for NS. However, this will not be successful and I will explain why. In my argument, the problem was that the microphysical state (S), on which Emily's deciding to raise her hand in that faculty meeting depends on, is nomologically closed. This means that the objective probability of S is identical to the probability of S solely given prior microphysical states and the laws of nature. I argued that this objective probability of S entailed that any conceivable ability Emily could have that would mentally causally impact S would have to be nomologically invisible. But, if it is nomologically invisible, it will not be able to make a difference (counterfactually) to whether or not S obtains. Thus, the conclusion of my argument (i.e., there is nothing Emily can do that would falsify S) is true by definition on List's view.

So, while List's view of mental causation is a brilliant way to avoid the problem Kim was worried about, it is perfectly consistent with my argument for the Fixity of the Microphysical Configuration.

4.3 Jessica Wilson's Subset View of Mental Causation

Christian List (2019) argued that causation is always intra-level but never inter-level (thus mental causation is only at the high macrophysical level and never reaches into fundamental microphysics). This sort of response is a way to protect mental causation in some sense but does

not negate the conclusion of my argument (that Emily cannot change or make the microphysical states her decisions depend on not obtain). Now it is time to consider another potential type of non-reductive physicalist view of mental causation. Several NRP theorists have discussed the relationship between physicalism, mental causation, and overdeterminism (i.e., Bennett, 2007; Loewer, 2002; 2011; 2015). However, few have received more recent attention than Jessica Wilson (2011; 2021). In this section, I am going to explain Wilson's *subset view of mental causation*, and how it potentially undermines my argument (by making an acceptable or coherent model of overdeterminism applicable for the Emily example).

In her recent book *Metaphysical Emergence* (2021), Jessica Wilson presents two forms of what she calls emergence: strong and weak. The "strong" version of emergentism is what most philosophers of mind think of when they speak of emergence, a form of non-physicalism. The "weak" version of emergentism is what I have referred to as non-reductive physicalism. As Wilson describes it, if weak emergentism is true, "one can have one's physicalism and one's distinctive higher-level efficacy, too" (2021, 82).

Wilson maintains that the non-reductive physicalists can allow that Emily has an ability to cause $\neg S$, and that this kind of mental causation would be a case of overdeterminism, "but maintain that the overdetermination here is of an unproblematic double-rock-throw variety" (Wilson, 2021, 44). So, mental causation, on their view, is simply an example of plausible overdeterminism like Jack and Sally's rock throwings mentioned earlier.

The way Wilson thinks about overdeterminism in cases like Emily is to suggest that Emily's mental features (events, states, properties, processes) must have a "non-empty proper subset of the token powers had by [the microphysical states on which the mental features depend on]" (Wilson,

2021, 75). So, in any given circumstances, Emily's microphysical state would have a range of causal powers and her mental state would have a selection of those causal powers.

A "non-empty proper subset" is a reference to a particular condition that Wilson coined (2011), which she calls the *Subset Condition of Causal Powers* (SCCP). She defines SCCP the following way:

Subset Condition on Causal Powers (SCCP): The token powers of a realized mental feature M on a given occasion are a non-empty proper subset of the token powers of the physical feature P realizing M on that occasion. (Wilson 2011: §1.2; see also Wilson 1999: 45)

In order to explain this theory, it may be helpful to elaborate on this notion more generally.

In this case, we are discussing the sets of causal abilities or powers of particular states or events. For any state or event, we could represent its causal profile as P (p1, p2, p3, p4). Generally, P would be a subset of P* simply in the case that P is contained within P*. If P and P* were identical, one could still say that P is a subset of P* and P* is a subset of P. A proper subset (like the one Wilson utilizes) is a subset relationship that is asymmetrical, where two sets are non-identical. Take the following illustration, let A stand for the set of all humans (where each constituent is a particular human being like George Washington, Winston Churchill, Frederick Douglass, etc.), and B stand for the set of all mammals. We can say A is a proper subset of B because all of the constituents of A (human beings) are constituents of B (mammals), but B's set is much larger than A's. A is not identical to, but is a proper subset of, B. The adjectival phrase "non-empty" is the only aspect of Wilson's SCCP theory that I have not discussed yet. Sets, technically, can be empty. Empty sets are technically subsets of all sets, so Wilson wants to be clear that the causal powers

of mental features are not ‘empty’ in this sense. This may seem like an unnecessary distinction, but it is technically necessary for Wilson. It also hedges against her view being compatible with epiphenomenalism. I will bring this up again in S 4.4.

Applying the SCCP condition to mental causation, Wilson argues that mental causation is a proper subset of microphysical causation. Mental events (M’s) have causal power profiles that are not identical to but rather proper subsets of microphysical events’ (P’s) causal power profiles. This is important because M’s causal power set is *distinctive*, and thus counterfactually tracks differently than P’s. For example, take something like Emily’s hand actually raising in the air and think about Emily’s intention to raise her hand in the air. In this case, S (the microphysical realizer of Emily’s intention or decision) could have been slightly different and still realized Emily’s intention (because her intention is multiply realizable). Thus, as Wilson and Bernstein point out, “the only powers that matter for the production of [Emily’s hand raising] are the powers associated with [Emily’s intention]; powers differing between [S*] and [S] (e.g., to produce different readings on a neuron detector) are irrelevant for the production of [Emily’s hand raising]” (Bernstein & Wilson, 2016, p. 321).

This is similar to List and Shoemaker’s difference-making approach from earlier, but there is a key distinction. Wilson (2021, p. 81) explains what the difference is in the following passage:

But here, in my view, Shoemaker’s discussion goes awry. I agree with Ney that the supposition that a Weakly emergent feature might be appropriately deemed ‘the’ cause (or a ‘more direct’ cause) of a given effect is problematic: she is right that it doesn’t make sense to deny that the lower-level realizer is a cause (and moreover a direct cause) of the effect, given that the supposition that the Proper Subset of Powers Condition is satisfied entails that the realizer has the power to produce (or contribute to producing) the effect,

and this power is manifested (or the associated regularity instanced, etc.) on the occasion in question. Moreover, as I note in Wilson 1999, 2011b, and elsewhere, the supposition that a higher-level feature causes something that its base feature does not cause is, for reasons that Kim's causal overdetermination concern brings out, in tension with physicalism and the associated supposition of Physical Causal Closure, whether the purported effect is a higher-level special-science feature or a lower-level physically acceptable feature. There is, however, no need to follow Shoemaker in suggesting that realized features can cause effects that their realizing features don't cause, since what the Weak emergentist requires is not that higher-level features be uniquely efficacious but just that they be distinctively efficacious.

The key part of this quote is the admission that on her view both the realizer and realizee are causes of the macrophysical effect. Think about Emily's hand-raising again (in the Faculty meeting). According to Wilson, we can suggest that her intention to raise her hand was caused by some previous mental state like her desire to contribute to the team, and this causal explanation would be because that desire made a difference to her hand-raising (in comparison to the microphysical state realizing her prior desire). However, Wilson is further suggesting that the microphysical state realizing Emily's prior desire could still be causally relevant (counterfactually, in a broader sense than List's difference-making analysis) to Emily's intention to raise her hand. One way of thinking about this is to say that M and P share some of the same causal oomph required for her hand-raising. Another way to explain this would be to suggest that there are two different kinds of causation (Cartwright, 2022). Regardless, in the case just mentioned, Emily's prior mental state (her desire) would be causing Emily's intention in the difference-making sense and the microphysical realizer of her desire would be counterfactually and causally relevant.

Of course, the danger in thinking about both the realizer and realizee as genuine causes of some event is the question of redundancy or reduction. If both the realizer and realizee *cause* the effect in question (even if they cause it in two senses), why not simply identify one with the other (reduce the mental to the physical)? This is Kim's worry about such a subset view (although Kim is talking about Terence Horgan's version of the view, which is similar to Wilson's). Kim (2010, p. 262) states this worry in the following way:

So, on Horgan's antireductionist view, a desire (as an individual "token" event) has "genuine" causal efficacy, say the power to move my arm to reach for a glass of water, in virtue of the fact that its neural realizer, an instance of a neural property on which the desire supervenes, has this causal power. Horgan's suggestion, I believe, is fundamentally right: mental events and states have the causal efficacy that they have because their neural/physical realizers have causal efficacy. In fact, a mental state, occurring on a given occasion, in virtue of being realized by a certain neural/physical state, has exactly the causal powers of that physical state. Where I differ from Horgan is that once we are prepared to say what we have just said, the next natural step to take-in my view, a step we are compelled to take- is to reductively identify this particular mental state with its neural-physical realizer. This of course is to jettison the "nonreductive" part of nonreductive physicalism.

Here, Kim is claiming that, if a particular mental event's causal power in a token event *is just* the causal power of the microphysical realizer, then it seems odd to suggest either that the mental event is really distinct or that its causal power is distinct.

To combat this fear of reductionism, Wilson employs the SCCP (proper subset condition). Recall that, according to Wilson's SCCP, anytime a mental event M occurs it has a distinctive subset of causal powers in relation to M's microphysical realizer's set of causal powers. So, while M's set

has no individual power that is distinctive from P's set, the set itself is distinctive and guarantees the difference-making causal relation (with macrophysical effects at least). As Wilson explains, "[t]his distinctive efficacy is not (even for Shoemaker, notwithstanding certain of his remarks' suggesting otherwise) a matter of a higher-level features being able to cause things that its realizer(s) cannot cause; rather, it is a matter of its profile's being such as to track relevant difference-making considerations, or comparatively abstract systems of laws or causal joints" (2021, 82).

To recap, a proper subset is a mathematical set that is asymmetrically included in another mathematical set ((1, 2, 3) is a proper subset of (1, 2, 3, 4, 5)). By saying that a mental event or property's causal profile is a non-empty proper-subset of its realizers causal profile is to say that (where M is a mental event and P is the corresponding realizer) all of M's causal powers are found in P, but the set of M's causal powers is distinct from P's (i.e., M (p1, p2, p3) is a proper subset of P (p1, p2, p3, p4, p5)). Because Wilson (and virtually all NRP theorists) assume a counterfactual or difference making view of causation, the distinct causal set of M is enough to ensure that M will be explanatory and causally distinct from P.

Wilson's subset view is a way to justify the overdeterminism necessary for non-reductive physicalism. Wilson's SCCP certainly seems like a plausible view, and thus, plausibly allows the NRP theorist to reject NS (The Fixity of the Microphysical Configuration).

Wilson's subset view of mental causation is ingenious and creative. However, I do not think it is the solution some might think it is. In this next section, I will argue that the subset view of mental causation ultimately fails.

4.4 The Problem of Additivity Again

In this section, I will argue against the subset view of mental causation explained in the previous section. In order to do this, I will begin by making a brief observation about a general problem with counterfactual causation. Following Paul (2007; 2015), I will then explain my objection to Wilson's subset view using a new Federer (and Ghost) case.

4.4.1 A General Problem with Subset Views

The Subset View of Mental Causation assumes a counterfactual notion of causation. Counterfactual theories of causation are any theory that explains the meaning of causal claims in terms of counterfactual conditions (e.g., "If the kettle was not hot, it would not have exploded"). With regards to counterfactual causation, critics have generally argued that the predominant varieties of counterfactual theories of causation entail that a great number of pseudo-causal cases (i.e., cases where correlative influence is present but not causal) are genuine causal cases, which is not desirable. In other words, the counterfactual view is sometimes argued to be under-sensitive when compared to our causal intuitions. Menzies and Beebe (Menzies & Beebe, 2020, p. S.3) illustrate this with an example originally from Bennett where rain influences a forest fire by delaying it:

The theory implies that any event that influences another event to a certain degree counts as one of its causes. But common sense is more discriminating about causes. To take an example of Jonathan Bennett (1987): rain in December delays a forest fire; if there had been no December rain, the forest would have caught fire in January rather than when it actually did in February. The rain influences the fire with respect to its timing, location, rapidity, and so forth. But common sense denies that the rain was a cause of *the fire*, though it allows that it is a cause of the *delay in the fire*.

Regarding Bennett's particular example, it would be very counterintuitive to suggest that rain caused the fire, but it does seem true to suggest that *if the rain had not occurred*, then the fire would not have occurred in February. In this simple example, what we learn is that counterfactual theories of causation may unknowingly track such non-causal events as being causal-events because counterfactual theories have a tendency to be under-sensitive.⁵⁹

In saying this, I am not suggesting that a tendency to be under-sensitive is a reason to wholesale reject counterfactual theories of causation. Counterfactual family of theories have unique problems to respond to and that, in of itself, is not a reason to reject the theory. However, I think that the counterfactual theory of causation has a substantial problem accounting for mental causation, and this problem is related to its tendency for under-sensitivity.

4.4.2 The Problem

Following Jaegwon Kim (1998; 2005; 2010) and L.A. Paul (2007; 2012), I will argue that accounting for mental causation in terms of a counterfactual analysis is uniquely problematic.⁶⁰ The subset view (i.e., the SCCP) grants mental properties' causal efficacy solely in virtue of distinct counterfactual relations between the given mental features and their microphysical realizers' effects. In the case of Emily (again), her intention to raise her hand is a mental state that

⁵⁹ Other general problems are cases of late and early preemption (see Paul, 2010).

⁶⁰ The argument tends to claim that epiphenomenalism is compatible with counterfactual theories of causation, thereby granting, paradoxically, epiphenomenal creatures mental causation (see Kim, 1998, 71; 2010, 254). I actually think this is a good argument, but there is a challenge here. The challenge is that to say a creature C is epiphenomenal *is to* say that it cannot mentally cause some effect E. Dialectically, this appears like question begging to some, so I am opting for a different strategy following Paul's problem of additivity mentioned in S.4.1.

depends on some microphysical state *S*. Prior to this moment, if we assume that Emily could have made *S* not obtain via the exercising of some ability, then, in this alternative sequence, we will be picturing Emily's prior mental state (e.g., a desire to not make herself seem foolish) and its microphysical realizer (i.e., *S**) both contributing to *S* not obtaining. However, *S** will be the difference-maker for *S* not obtaining because it is at the same metaphysical layer of the world. Emily's prior mental state (e.g., her desire to not make herself seem foolish) will be counterfactually related in a more indirect sense to *S*'s not obtaining than *S**'s causal contribution. I shall argue that the asymmetry between the mental features and the microphysical features (which the mental features depend on) proves that the subset view of mental causation is not an acceptable theory of mental causation.⁶¹

To explain this argument, I am going to refer back to the Federer case from § 4.1.⁶² Then, I am going to offer other versions of the case to make the argument that Emily's mental features are in the same counterfactual *and* causal position as the Ghost, which enables me also to argue that Emily's mental features cannot be genuinely causal on the subset view.

In §4.1, I discussed the Federer Case. The Federer case is notably a case where

⁶¹ Hiddleston (see Wilson, 2011, S. 2.3) and W. Robinson (2019) seem to indicate that they think the SCCP is straightforwardly compatible with epiphenomenalism. This might be true, but I think it is really dependent on one's theory of causation. If one opts for a causal power view, I cannot see why the SCCP would be straightforwardly compatible with epiphenomenalism.

⁶² That case stated: **Federer**: Roger Federer, the tennis player, hits a tennis ball with his racket at *t*, transferring momentum (*p*) to the ball, changing its direction, and flinging it across the net. Unbeknownst to Federer (but seeable to the crowd), a ghost appears right before Federer strikes the ball and strikes the ball with an ethereal tennis racket at *t*. However, the tennis ball exhibits the same amount of momentum as it would have if only Federer had struck it.

- (I) Federer's strike and the Ghost's strike at t are supposed to *overdetermine* the ball's flying over the net at t_1 .

However, the following is also true

- (II) The tennis ball is asymmetrically nomologically related to Federer's strike.

By (I), I mean that there is an appearance of overdeterminism in the case, but not necessarily a genuine occurrence of it. I previously suggested that the natural conclusion to draw from the case is to believe that the ghost did not really cause the ball to fly over the net at t (it is a case of *apparent* but not *genuine* overdeterminism). There are two reasons for this. Firstly, it seems clear that if the Ghost had not appeared, the ball would have still sailed over the net at t_1 . But, if Federer did not strike the ball, the ball would not have sailed over the net at t_1 . Secondly, if an observer saw this situation, it would be an *open* question to them whether the ghost did cause the ball to fly over the net at t_1 . Given that the physical features of the case are such that the Ghost's strike did not *add* anything nomologically relevant to the ball, the simpler explanation would be to say the Ghost did not genuinely cause the ball to fly over the net at t_1 .

The Federer case *appears* to be overdeterminism, but the asymmetry casts doubt on this. Thus, the inference should be that it is not genuinely overdeterminism and the ghost is not causally doing anything efficacious in this scenario.

However, what if we slightly adjusted this case? Consider the following amended version:

Federer*: Roger Federer, the tennis player, hits a tennis ball with his racket at t , transferring momentum (p) to the ball, changing its direction, and flinging it across the net. Unbeknownst to Federer (but seeable to the crowd), a ghost appears right before Federer

strikes the ball and strikes the ball with an ethereal tennis racket at t . This ghost loves Federer and wants to mimic him, and thus, has chosen to appear *every time Federer is about to hit the ball in order to smack the ball with Federer*. However, the tennis ball exhibits the same amount of momentum as it would have if only Federer had struck it.

In this case, there is a pattern between Federer's and Ghost's striking of the ball. If the Ghost's choice to mimic Federer is consistent enough, then presumably, the following counterfactual would be true: if the Ghost had not struck the ball at t , the ball would not have flown over the net at t_1 . This is true because, in the closest possible worlds where the Ghost does not strike the ball, neither does Federer (and as a result, the ball does not fly over the net).⁶³

So, here the counterfactual relation between Federer's hitting the ball and the ball flying over the net is symmetrical with the Ghost's hitting the ball and the ball flying over the net. However, the additivity is not symmetrical. Federer's hitting the ball can add to the nomological probability, while the Ghost's hitting cannot. Should we now infer that the Ghost did, in fact, genuinely overdetermine the tennis ball flying over the net t_1 ?

The answer here should be negative. For, if the correct causal inference in Federer is that the Ghost did not genuinely cause the tennis ball to fly over the net at t_1 , then the correct causal inference in Federer* too should be that the Ghost did not genuinely cause the tennis ball to fly over the net t_1 .

⁶³ If one does not think this is a strong enough link, then we can simply amend the case. Imagine, instead of the Ghost simply liking Federer, what if there is a curse on Federer such that the Ghost, by magical law, is always appearing in situations like this.

Why? Because there is no relevant difference between the Ghost's action in these cases. The only difference is how frequently the Ghost appears (and tricks observers).

My argument thus is the following:

- (I) In Federer, the Ghost is not causing the ball to fly over the net at t_1 because of the asymmetrical nomological relation to Federer's strike at t .
- (II) There is no relevant difference between what the Ghost does in Federer to what he does in Federer*.
- (III) Therefore, in Federer*, the Ghost is not causing the ball to fly over the net t_1 .

Of course, here, the problem is that by applying the subset view to Emily (and Emily raising her hand at the faculty meeting), her potential decision is in the same position as the Ghost (in Federer*). Thus, the counterfactuals between the microphysical state and mental states would be symmetrical, but the additivity (given NOMOLOGICAL CLOSURE) would be asymmetrical.

Thus, we can infer:

- (IV) There is no relevant difference about Emily's mental features and the Ghost's mental features in Federer*.
- (V) Therefore, Emily, via her mental features, could not cause $\neg S$.

If Emily, via her mental features, could not causally make S not obtain, then NS is going to be true. The only manner to do so would theoretically be to weaken mental causation to a mere counterfactual relation, but there are too many potential counterexamples (e.g., the Ghost) that would make this move implausible for mental causation. The goal for this chapter was to give an

argument for the FIXITY OF THE MICROPHYSICAL CONFIGURATION (NS), which is what this argument effectively concludes.

Conclusion

I argued in the last section was that in the Federer and Ghost case there was a genuine case of overdeterminism such that the following are true:

- (1) If Federer had not struck the ball, the ball would not have flown over the net.
- (2) If the Ghost had not struck the ball, the ball would not have flown over the net.

However, even though (1) and (2) are plausibly true, in this scenario, Federer's striking the ball *adds* to the objective nomological probability of the ball flying over the net, while the Ghost's striking the ball *does not add* to the objective nomological probability of the ball flying over the net. This should cast serious doubt on whether the following is true:

- (3) The Ghost's striking the ball overdeterminately *caused* the ball flying over the net.

In fact, I think the additivity difference should lead us to reject (3) altogether. In the traditional Federer case, there was not a counterfactual like (2), and the intuitive inference would be that the Ghost did not really cause the ball flying over the net. If we alter the scenario in Federer* to get (2) to be true, the Ghost *does not do anything different than he did in the original scenario*. Therefore, the Ghost should not be counted as a genuine overdeterminately cause of the ball's flying over the net *even if* (2) holds true. With regard's the Emily raising her hand the in the non-reductive physical world, Emily's mental state (i.e., her decision) is in the same position as the Ghost with relation to $\neg S$ (where S is a microphysical state): Emily's mental state is not additive, but is counterfactually sufficient for $\neg S$. Thus, we Emily's mental state is not overdeterminately *causing* $\neg S$. And, thus, we can infer that NS is true for Emily in a non-reductive physicalist world.

And arguing for the truthhood of NS (THE FIXITY OF THE MICROPHYSICAL CONFIGURATION) was the goal of this chapter.

CHAPTER FIVE. ALTERNATIVE POSSIBILITIES

Introduction

In Chapters Two and Three, I explained two versions of the modal consequence argument: The P-MODAL CONSEQUENCE ARGUMENT and the B-MODAL CONSEQUENCE ARGUMENT. The P-MODAL CONSEQUENCE ARGUMENT is an argument concluding that, necessarily, determinism and free will are incompatible. The B-MODAL CONSEQUENCE ARGUMENT concludes that, necessarily, non-reductive physicalism and free will are incompatible. The main focus of this thesis is on the B-MODAL CONSEQUENCE ARGUMENT. The two main avenues for criticism against the B-MODAL CONSEQUENCE ARGUMENT are the Beta modal inference rule (common to both the B and P-MODAL CONSEQUENCE ARGUMENTS) and the FIXITY OF THE MICROPHYSICAL CONFIGURATION principle (exclusive to the B-MODAL CONSEQUENCE ARGUMENT). I discussed why some critics of the modal consequence argument (i.e., the B and P-MODAL CONSEQUENCE ARGUMENTS) are skeptical of the Beta modal inference rule in Chapter Two (§ 2.2-2.4). In Chapter Four, I defended the FIXITY OF THE MICROPHYSICAL CONFIGURATION principle. So, it is true that I have supported the B-MODAL CONSEQUENCE ARGUMENT against its most common criticisms.

However, there is another sort of objection that needs to be discussed now. Some critics think the critical problem with all versions of the consequence argument is related to the assumption of ALTERNATIVE-POSSIBILITIES (which states a creature *C* is free in ϕ -ing *only if* *C*, when choosing to ϕ , could have decided to $\neg \phi$). In the context of the B-MODAL CONSEQUENCE ARGUMENT, this principle is used to infer the following in premise four of the argument:

- (4) Necessarily, if there is nothing that Emily (or anyone) can do that would make Emily's deciding to raise her hand at t not obtain, then Emily cannot freely decide to raise her hand at t . (Assumption from ALTERNATIVE-POSSIBILITIES).

I take it that ALTERNATIVE-POSSIBILITIES is the third most controversial element of the B-MODAL CONSEQUENCE ARGUMENT, and, as such, must be addressed. In order to address potential criticisms, I will first explain a modal interpretation of ALTERNATIVE-POSSIBILITIES following John Maier (2015; 2020; 2022). By explaining Maier's interpretation (or a potential interpretation using Maier's notion of ability), we can explain *why* and *how* alternative possibilities are metaphysically necessary for free will.

I will subsequently defend the B-MODAL CONSEQUENCE ARGUMENT against two objections related to ALTERNATIVE-POSSIBILITIES. The first objection comes from Maier (2022), who argues that on a modal view of abilities agents can have relevant alternative possibilities in deterministic and physicalistic worlds. Although, Maier's criticism is not directly against my version of the consequence argument, it can be interpreted as an objection related to alternative possibilities. The second objection concerns the so-called Frankfurt cases. In these cases, agents appear to be acting freely *without alternative possibilities*. The Frankfurt-styled objection to the consequence argument can grant that the FIXITY OF THE MICROPHYSICAL CONFIGURATION principle is true (so *Premise 2 is true*) and the Beta Rule is valid (so *Premise 3 is true*), and yet argue that the consequence argument (generally) does not entail Emily cannot *freely* decide to raise her hand at

t (*Premise 5 is false*). The only premise between Premise 3 and Premise 5 is Premise 4, which is an expression of ALTERNATIVE-POSSIBILITIES.⁶⁴

This chapter is divided into two main parts: The first half discusses Maier, and the second half delves into Frankfurt-styled Arguments. In the first half (§5.1), I will begin by outlining the modal theory of abilities, and John Maier’s particular theory of abilities. Next, I will address a potential question known as the Robustness Objection (§5.1.3), and demonstrate how Maier’s view can respond (§5.1.4). Furthermore, I will elucidate Maier’s objection (§5.1.5) to the consequence argument, which pertains to all versions (including, potentially, my own), and offer my own response to Maier (§5.1.6). Finally, I will consider two possible ways to further interpret Maier’s views of options (§5.1.7).

In the second half (§5.2), I will discuss the famous Frankfurt cases and the arguments they are assumed to support. In this section, I will consider four different varieties of proposed Frankfurt-styled Arguments and argue that none offer a convincing objection to the ALTERNATIVE-POSSIBILITIES or the B-MODAL CONSEQUENCE ARGUMENT.

5.1 John Maier’s Modal Theory of Abilities

The goal of this section is to explain John Maier’s modal theory of abilities. Ultimately, I am going to propose that we interpret ALTERNATIVE-POSSIBILITIES through Maier’s concept of *options*,

⁶⁴ Although, in Maier’s case, he might elect to deny either the fixity principle or Beta. It is slightly difficult to locate which premise/inference rule he would reject because he rejects a premise in a different form of the argument. It is clear though that he thinks *something* is wrong with the argument. His criticism ends up being similar, in certain respects, to David Lewis’s (1981), Kadri Vivhelin’s (2013), Cian Dorr’s (2019), and Christian List’s (2019).

which I will explain this section (§ 5.1.1-5.1.4). Additionally, I will explain Maier’s criticism of the consequence argument and offer a response to that argument (§ 5.1.5-5.1.6). Finally, I will offer a possible augmentation of Maier’s view (§ 5.1.7).

5.1.1 Modal Abilities

In a typical choice situation, a conversation like the following can occur:

Haddie: I bought strawberry ice cream, what kind of ice cream did you buy?

Eila: Oh, I could have purchased strawberry, I thought about it, but I ended up buying vanilla.

These types of claims are reflective of our abilities. Eila claims she could have bought strawberry ice cream but chose to go with vanilla instead. There are different ways of conceptualizing abilities, but one broad umbrella is what Maier (2020) calls the *modal theory of ability*.⁶⁵ The modal theory understands all abilities in terms of some kind of possibility claims. A modal theory of ability is, as Maier, defines it, “any theory on which claims about an agent’s abilities are understood in terms of claims about what that agent does at some possible world (or set of worlds)” (Maier, 2020). There are different ways of constructing a metaphysically robust modal theory of abilities. Still, all modal theories generally want to give some kind of account or analysis of what it is to have an ability in terms of some subset of what is possible.

⁶⁵ For a helpful overview of other views of abilities, see Maier (2020).

5.1.2 Options and Abilities

I am going to begin by explaining a modal theory of abilities from John Maier (2015; 2020; 2022). According to Maier's view, all subset categories of abilities are analyzed in terms of something he calls *options*. In an influential paper, Maier (2015) proposed a new kind of practical modality called an *agentive modality*. This agentive modality is supposed to be a way to capture an important aspect of agency that is different from other practical modalities. To explain this idea, it will be helpful to consider a locution like

(I) "Ben *can* play hockey."

This locution can be read in a couple of different ways.⁶⁶ The kind of theory Maier advocates, following Kratzer's semantics (1977), assumes that (I) is a possibility statement, meaning that "Ben can play hockey" is true *iff* there is a possible world where Ben is playing hockey. If one wished to assess the truth value of (I), one naturally would try to determine what Ben's relevant properties are that could make (I)'s possibility plausible. And, which of Ben's properties would be relevant to (I)'s truth value would depend on the context in which (I) is uttered.

To illustrate this last point, let us imagine that Emily wants her husband Ben to play hockey because Ben enjoys it (and she is a fantastic wife), but, alas, Ben cannot do so because there is no

⁶⁶ Notably as either properties (predicates) as Vetter (2010) proposes, or as possibility statements. Maier admits he is tempted to combine his account with Vetter's but for the purposes of this paper I am referencing, he assumes they are possibility statements like Kratzer (1977) does. I think there is a rich potential to take the work of Vetter and Maier together to form a more complete Agentive Modality that does not use possible worlds. However, this would be for exploration in a further project.

ice rink near them.⁶⁷ Emily might then utter “Ben *can* play hockey (!)” upon learning that there now is a recently constructed ice rink down the street. Here, Ben’s property of being within a certain distance of an ice rink makes (I) true because that is the relevant fact Emily cares about in the context in which she utters (I). Ben playing hockey is compossible with the fact that there is an ice rink within a certain distance. So (1) can be referencing this set of possible worlds. In can-statements (e.g., “Ben can play hockey”), persons care about different things, which makes a very diverse set of conditions for what makes can-statements true. So, this would be one way of reading (1) modally.

However, Maier argues that statements like (I) can be read as something *more* than just expressing a simple (Kratzerian) modality, as expressing something he calls an *agentive modality*. In our discussion of how to read (1) above, we implicitly assumed that all relevant facts (relevant for (1)’s truthfulness) are essentially non-agential. For example, it was Ben’s distance to an ice rink that was the relevant fact in the example above. A distance relation is not an agential fact (or relation). By contrast, when we talk about agentive modalities, according to Maier’s account, we are assuming we care about relevant facts that are *primitively* agentive in some respect.⁶⁸

⁶⁷ This is unfortunately has more truth than I wish it did.

⁶⁸ This account has been very influential since its introduction. Christian List, whom I have discussed repeating in this thesis, assumes that Maier is essentially right. He says, “[c]onsequently, if we wish to talk about agents and their actions, we must switch to a language of psychology, specifically one in which concepts pertaining to intentional agency can be expressed, together with the relevant modal notions: [Maier’s] agential ‘can’” (List, 2019, p. 257).

To better explain this account, I will use Maier's own example, the locution "Rachel can speak Mandarin." On Kratzer's account, the truth of "Rachel can speak Mandarin" will depend on certain facts concerning Rachel and her world. For example, contextually someone might be wondering whether or not Rachel has a gag in her mouth. And, if she does not, then Rachel *can* speak Mandarin. Alternatively, Maier notes that we could be concerned with some different kinds of facts than if she has a gag in her mouth. For example, "[w]e may be asking what she is *now able* to do, what she is *free* to do, what acts are *available* to her, or what is in *her power* to do" (Maier 2015: 119-120). These are modal facts that Maier calls *agentive*. Arguably, these sorts of questions (e.g., what is in Rachel's power to do now?) are not naturally reducible to non-agentive modalities. These are natural questions about the agent's powers, abilities, dispositions, or other properties. Maier thus believes that these particular questions "pick out the agentive modality on which I will focus here: they pick out Rachel's **options**" (Maier 2015: 119-120). Options are the foundational piece of Maier's system. Options are supposed to be a *primitive* or *fundamental* modal notion that does not reduce to other kinds of modal facts.

So, what are options (in Maier's conception)? According to Maier, options are the objects that persons deliberate about in common choice situations (2015, p. 120). They are intuitive and context-sensitive (at least to some degree). We might ask, what options do agents have in a typical situation? Maier responds (2022, 135):

I have presupposed that agents typically face a plurality of options in any given situation, and that their options correspond more or less to the options that we typically take them to have. I have presupposed, for instance, that someone with some available money in a shop that sells newspapers has the option of buying a newspaper, whether or not she in fact buys

one. I think this is how things are generally with options. There are no general rules that decide whether someone has an option in any given case.

With this newspaper shop example, we can observe that the relevant options are supposed to be the intuitive options that agents believe they have (whether agents are actually aware of them or not in a given situation). In choice situations, we intuitively believe that we have multiple options, so we assume that we have numerous options in any given situation (Maier, 2022, p. 136).

Although Maier does mention (in that quotation above) there are not general rules concerning how many options an agent has in a given case, Maier (2022, 135) does provide two general modal rules that we can use to connect options with other modal rules (and here OPT stands for “has an option”):

Performance Principle: If S performs an act of type *a*, then S OPT *a*.

Possibility Principle: If S OPT *a*, then there is a possible world at which S performs an act of type *a*.

The Performance Principle is suggesting that with respect to any action I actually do, I have an option to do that action. Maier does not explicitly commit himself to the thesis that agents *necessarily* have a multitude of options. However, options are supposed to be the paths one *might* take in a given choice situation. One interpretation is that Maier thinks that agency requires an

agent to have alternative options.⁶⁹ However, one could interpret Maier as saying having alternative options is normative and, hence, sufficient for agency. Exactly which is the best interpretation of Maier is not particularly important.

When I perform an action *a*, I (at least) typically also have the option not to *a*. In the newspaper shop example that Maier gave (assuming the agent bought a newspaper), because the agent bought a newspaper, we can surmise from the Performance Principle that the agent had the option to buy a newspaper. Additionally, we intuitively want to suggest that the agent also had the option to refrain from buying the newspaper, and Maier (2022, 134) would call this option an *unexercised option*. He claims, after all, that “an agent has an *unexercised* option to *A* just in case *S* OPT *a* and *S* does not actually perform an act of type *a*” (ibid). Additionally, the Possibility Principle would entail that if I had an option to buy a newspaper (because I actually bought it), then there is a possible world where I buy the newspaper. On his view, there is nothing more formally to say about options than these two principles.

However, Maier does have more to say about the concept of an ability. Agents may have a general ability to perform an action that is not an option in a given situation. For example, I can have the ability to play hockey but no current option to play hockey (because there is no rink). Agents may also have an option to do something that she does not have an ability to do. Maier writes: “general abilities are robust, in the sense that whether an agent has a certain general ability depends on what

⁶⁹ Taylor Cyr (2017) takes this interpretation for example. I tend to think that is the natural reading of Maier, but not much turns on whether this reading of Maier is right.

she is like in a wide range of circumstances, not just on how she is at the present moment” (Maier, 2015, p. 125). He defines abilities in the following way:

Ability: S has the general ability to A iff S has, in a suitable proportion of similar circumstances, the option of Aing. (Maier 2015: 127)

This is compatible with the idea that Sally might not have the ability to speak Mandarin even if she can utter a phrase merely by guessing. Barbara Vetter (2018) has a similar view (although in her view, *potentialities* are used in place of *options*) and she uses the example of hitting a bull’s eye while throwing darts (222). It may be true that I *can* hit a bullseye in the sense that I have an option to do so (my arms are working and I can throw the dart at the board even if I am lacking the skill) even if this option does not qualify as an ability (because in similar circumstances I cannot make this happen (ibid)).⁷⁰

What is important though is that we are defining *ability* in terms of *options*.⁷¹ The main reason why Maier’s view has been important in the recent literature is that he has shown a path for a *unified* modal theory of agency, where agential modal notions do not reduce to non-agential factors. A lot of prior work on modal theories of abilities try to analyze abilities in terms of non-agential factors (e.g., one’s circumstances in practical modalities). If Maier is right about the interpretation of

⁷⁰ Although, the control condition here might make it difficult.

⁷¹ There is a similar move with other agentive aspects, but going through all of Maier’s distinctions is not needed for my point.

certain agential can-statements, then analyzing abilities requires more than just the task of fine-tuning modalities – it requires considering agential modalities.

5.1.3 The Robustness Objection

This section will explain Taylor Cyr's (2022) and John Martin Fischer's (1994) Robustness Objection. This objection primarily challenges the ALTERNATIVE-POSSIBILITIES proponent to clarify the possibilities necessary for free will.⁷²

We can explain this objection through the question: Exactly *how* and *why* do an agent's alternative possibilities (at least) partially explain the agent's freedom in the actual world? In other words, why does having alternative possibilities matter for free will in the first place? Fischer points out that many imaginable alternative possibilities have nothing to do with one's freedom. For example, I may be struck by lightning right now. Being potentially able to be struck by lightning, as an alternative possibility to what actually happens, seems irrelevant to one's free will. So, what is it about a certain subset of the alternative possibilities that makes that subset relevant for free will? Here is Fischer (1994, pp. 140-141) articulating this worry:

⁷² It is important to note that this objection is (I) typically concerning variations of the AP principle concerning moral responsibility (e.g., A create C is morally responsible *iff* C could have done otherwise), which is different than ALTERNATIVE-POSSIBILITIES and (II) it is typically discussed in the context of the Flicker Response to Frankfurt-styled Arguments. However, I think the objection can be applied more broadly to clarify ALTERNATIVE-POSSIBILITIES. The reason I bring this up *before* discussing Frankfurt cases is that if we spend time considering the nature of alternative possibilities relevant for free will, we have a clearer concept to consider in the Frankfurt cases. This is probably more of a judgement or dialectical preference, but that is why I am presenting this first. Thanks to Taylor Cyr (P.C.) for pointing out that some readers may be curious why I choose to discuss this response first.

[What] needs also to be shown is that these alternative possibilities play a specific role in the appropriate understanding of the cases. That is, it needs to be shown that these alternative possibilities ground our attributions of moral responsibility...The existence of various genuinely open pathways is alleged [by the ALTERNATIVE-POSSIBILITIES theorist] to be crucial to the idea that one has control of the relevant kind. But if this is so, I suggest that it would be very puzzling and unnatural to suppose that it is the existence of various alternative pathways along which the agent does not act freely that render it true that the agent has the relevant kind of control (regulative control).

Taylor Cyr (2022, pp. 23-24) elaborates further giving a helpful example,

But the mere possibility of something else happening to an agent in an alternative sequence is insufficient to ground the agent's moral responsibility in the actual sequence. To see this, consider the following scenario. Suppose that causal determinism precludes [Alternative-possibility] freedom, and suppose that an agent who is causally determined to perform a particular action satisfies all conditions on moral responsibility minus having [Alternative-possibility] freedom (if this is a condition on moral responsibility). Now suppose that the case is tweaked such that the agent's action is not causally determined, but suppose that the reason the action is not determined is because there is a slight but non-negligible chance that the agent will be incapacitated then. If the agent whose action is causally determined is not morally responsible for that action, then neither is the agent whose action is undetermined in the way just described. Because mere alternatives, construed as the possibility of something else happening to the agent, are not such that agents are free in virtue of having them, they are insufficient to ground freedom on the [Alternative-possibility] model.

The situation Cyr describes is illuminating in figuring out what this problem is supposed to be. In Cyr's scenario, in possible world w , an agent is not morally responsible (because the agent does not have alternative possibilities). However, in a different possible world w^* , which is precisely like w except that determinism is not *quite* true (i.e., something like almost-determinism is true) and, given that determinism is not quite true, it is possible that the agent is struck by lightning. Cyr is arguing that, in w^* , the agent would be able to do otherwise *in some sense*. But, as Cyr and Fischer try to point out, this particular alternative possibility does not seem like a good candidate to be part of the necessary conditions for an agent's freedom (or responsibility).⁷³

I take Cyr and Fischer to be making an objection that is principally pushing the theorist to clarify ALTERNATIVE-POSSIBILITIES. Two questions are easily distinguishable:

- I. Firstly, *why* (or *how*) are alternative possibilities metaphysically necessary for free will? In other words, what kind of "can" statement does one reference in statements like "he *can* do otherwise" and *why* should this sense of "can" be in the necessary condition for free will?
- II. Secondly, it is obvious that not all alternative possibilities are relevant, so *which* kinds of alternative possibilities are relevant? In other words, once we know a particular kind of a can-statement is being referenced (e.g., "he *can* do otherwise"), how does this help us restrict irrelevant alternative possibilities (e.g., the possibility of being struck by lightning)?

⁷³ I am more interested in the agent's free will than moral responsibility. Cyr is more interested in responsibility.

These are related questions, but they are different. Both need to be answered. If there is no way to adequately respond to (I) and (II), then the critic of the B-MODAL CONSEQUENCE ARGUMENT can suggest that we have no reason to accept that not having alternative possibilities entails not having free will (premise 4 of the B-MODAL CONSEQUENCE ARGUMENT).

5.1.4 Maier and the Robustness Objection

We can theoretically utilize Maier's notion of options and abilities to respond to the Robustness Objection- at least, the question of *why* or *how* alternative possibilities are necessary for free will. Let me explain this potential response next.

On Maier's account of ability, all abilities are dependent on options; in other words, options are the modal building blocks of agency. If an agent never had any options, this would entail the agent would never have any abilities or agency. The way I originally defined ALTERNATIVE-POSSIBILITIES was:

ALTERNATIVE-POSSIBILITIES: A creature *C* is free in ϕ -ing *only if* *C*, when choosing to ϕ , could decide to $\neg \phi$.

Putting this condition in the language and framework of Maier's agential modalities, we will have choice of interpreting "could decide to $\neg \phi$ " as either "had the option to $\neg \phi$ " or "had the ability to $\neg \phi$."⁷⁴ Because our normal experience in choice situations is that we have multiple options, the

⁷⁴ We could also interpret it as other kinds of agential categories like skills, but I think abilities and options are the most natural interpretations.

more natural way to utilize Maier's framework is the following revision of ALTERNATIVE-POSSIBILITIES:

OPTIONS-BASED ALTERNATIVE-POSSIBILITIES: A creature C is free in ϕ -ing *only if* C, when choosing to ϕ , had the option to decide to $\neg \phi$.

This would be effectively a kind of new rule connecting options to free will (one that Maier does not himself explicitly suggest).

In this definition, we are explicitly using Maier's notion of options to make clear the modality involved in the alternative possibilities relevant for free will. And, recall, on this modal theory of abilities, one cannot have abilities and skills if one does not have options. Thus, options are essential for agency. In any given choice scenario, without having an alternative option, one would not be free because one would not have the requisite abilities or skills. This seems like a sufficient response to the question of why alternatives would be important at all-the first prong of the Robustness Objection.

Recall, I separated two prongs of the Robustness Objection. The second prong is asking for an account that would be able to specify which kinds of alternative possibilities are relevant to free will. What makes certain alternative possibilities irrelevant to freedom and other ones to relevant to freedom and agency? In other words, what makes the possibility of being struck by lightning in a given choice scenario *not an option*?

On Maier's view, in order to genuinely act (which is where we get options from in the Performance Principle), we must have a good amount of *direct control* over the outcome of our action (e.g., drawing a card or drawing a spade). For example, right now I have a good deal of control over whether or not I drink my coffee. That is a genuine option of mine. I have little to no control over

whether I get struck by lightning. Therefore, it is not a genuine option. One helpful heuristic to think of control is probabilities. If I form an intention to drink my coffee, all things being equal, there is a high probability I will drink the coffee. However, regardless of my intention, the odds I get struck by lightning in my lifetime is something like .00006 (and much lower if we ask what is the odds I get struck *right now*). In this sense, I do not have the same kind of control over getting struck by lightning as I do drinking my coffee, thus I can pretty clearly suggest I have an option to drink my coffee, but not to get struck by lightning.

In the next section, I will explain how Maier's thinking of options has evolved and how his most recent comments on options (2022) undercut the consequence argument in an interesting way.

5.1.5 Maier on the Consequence Argument

One question that Maier has answered differently over the years is whether or not an agent an agent in a deterministic universe can have multiple options (e.g., the options she exercises and additional unexercised options)? In his (2015, pp. 132-133), Maier notes:

Can it be the case that A is an option for S though it is not possible, relative to the laws and the past, that S does A? If one answers this question in the affirmative, then one is a compatibilist in the relevant sense; if in the negative, then one is an incompatibilist. Nothing said thus far demands one answer or the other to this question.

Here, Maier clearly suggests that his account is open to be further developed affirmatively or negatively on this question. However, in his (2022), Maier claims that he thinks his notion of options implies that agents *do have alternative (unexercised) options in deterministic universes*. To show what changed in Maier's thinking, we can review Maier's take on his interpretation of Van Inwagen's consequence argument (1983).

Maier interprets Van Inwagen's consequence argument to be a counterfactual argument (similar to Lewis' interpretation, 1981). In Maier's reading of the consequence argument, Maier (2022, pg., 144) takes the key premise (stated in the language of options) to be:

(Option-Inheritance): If the agent had the option to raise his hand, and the conjunction of the past and laws entails that he did not raise his hand, then the agent had the option of rendering the past false or the option of rendering the laws false.

This is premise Maier ultimately objects to, but, before I discuss that, here is the rest of the argument as Maier presents it (2022, pg., 143):

(Past-Fixity) If the agent could have rendered the past and the laws false, then the agent could have rendered the laws false.

(Law-Fixity) The agent could not have rendered the laws false

[Conclusion 1]: It follows either that the agent could not have raised his hand or that the conjunction of the past and the laws does not entail that he did not raise his hand.

[Conclusion 2]: That implies that either agents lack unexercised options or determinism is false.

Conclusion 2 here is essentially stating that either having unexercised options is incompatible with determinism or determinism is false. (And, I assume, by "false" Maier means incoherent or metaphysically impossible.) Either disjunct (in Conclusion 2) would be favorable for Incompatibilism.

This version of the consequence would provide an argument that agents in deterministic (or physicalistic) worlds do not have alternative options. It is important for Maier to object to this

argument, so that he can maintain his stance that there are only two rules governing options. If this argument, or the P or B-MODAL CONSEQUENCE ARGUMENT are found to be sound and convincing, then we can naturally add to this set of rules governing options.

To be clear, the Option-Inheritance premise is not a premise in my version of either the P or B MODAL CONSEQUENCE ARGUMENT. However, I think there is a way to interpret Maier's of Option-Inheritance that is relevant for the B- MODAL CONSEQUENCE ARGUMENT.⁷⁵ I take it that Maier would argue something akin to the argument that a proper understanding of alternative possibilities (i.e., options) shows that agents *can be free* in deterministic (and physicalistic) worlds, and thus, by reductio, we should reject either the fixity principle of BETA-BEST. However, in Maier's retelling of the consequence argument, Option-Inheritance is a premise, and Maier denies this premise for two separate reasons.

Option-Inheritance is a premise expressing that agents, who theoretically have multiple options in a deterministic world, would additionally have an option of changing the past or the laws.⁷⁶ On the face of it, suggesting ordinary humans have an option to change the past or laws seems ridiculous because we would be impugning ordinary human agents with superhero-like or divine powers.⁷⁷

⁷⁵ And the P- MODAL CONSEQUENCE ARGUMENT.

⁷⁶ Here, he using the language of falsifying because he is thinking of the past as a proposition. We can think of these as states as well and then suggest that one has the option of making the state of the completed past not obtain or a different state obtain.

⁷⁷ And, it might even be more than that. On many conceptions of divine omnipotence, God does not have the power to change the past (although he can change the laws). Barry Loewer, in fact, does suggest that ordinary humans do have the option to change the past, just in an insignificant way, which is his way of rejecting the consequence argument, see Loewer (Forthcoming). Dorr (2016), similarly, hints at such an approach. It seems like the conditional consequence of being able to change the past is a good place to argue a reductio against their broader assumptions.

Maier's first explanation for why Option-Inheritance is false is that "it does not follow from either of our two principles about options" (2022, 144). It does not follow from the Performance or Possibility Principles directly. The Possibility Principle states that if *a* is an option for P, then there is a possible world where P performs an action of type *a*. There is no clear way to reason from that principle to Option-Inheritance. The Performance Principle states that if a person P performs an action of type *a*, then P had an option to *a*. Maier is right that one cannot logically reason from these principles to Option-Inheritance. Because Option-Inheritance does not follow, Maier argues that given his modal theory of abilities, the proponent of the consequence argument must be implicitly adding in some further principle that is effectively question-begging.⁷⁸

However, I do not think question-begging is a very persuasive response: Option-Inheritance does not follow from the Possibility or Performance Principles because, according to Maier (2022), these are the *only* rules governing options. However, Maier does not provide a reason as to why we should simply stop at these two rules.⁷⁹ The obvious response to this argument is to suggest that we know there are more than two rules governing options because of the consequence argument. Getting more rules requires further argumentation regarding options *like the consequence argument*. In fact, if virtually any version of the consequence argument (e.g., the P or B-MODAL CONSEQUENCE ARGUMENT) is sound, then Option-Inheritance becomes a

⁷⁸ Question-begging has been a frequent criticism as of late for the consequence argument. In her recent debate book with Robert Kane, Carolina Sartorio also argued that the consequence argument cannot succeed without question-begging (Sartorio & Kane, 2022). You can interpret Lewis as making a similar criticism in his famous paper (Lewis, 1981) as well. However, neither Lewis nor Sartorio interacts with the version of the consequence argument I mustered in this thesis.

⁷⁹ Sometimes Maier does mention that options are primitive and therefore resistant toward analysis, but he still owes an explanation for *his* analysis.

counterpossible because the antecedent would assume something impossible; that is, it would assume agents would have unexercised options in deterministic worlds.

Maier does, however, also offer a second response to Option-Inheritance that is more compelling. He explains that we cannot have the option to falsify the past or the laws by the nature of options themselves (2022, 145):

Generally speaking, our options are in a certain respect inherently coarse-grained. *I may have the option of picking a card from the deck and yet not have the option, for any particular card, of picking that card. Our control over the world is in this respect imperfect.* Imagine what things would be like if this were not so. An agent who was not imperfect would have the option of performing specific forms of any act she might perform, for arbitrary degrees of precision. She would have available to her a vast range of acts that are not available to creatures like us. She would have the option of making exceedingly fine-grained interventions in the world. These acts would at least be the right kind of acts to break the laws. We are not like this. *In a deterministic world, our unexercised options are such that, if we perform them, the laws or the past would have to be different. But we do not actually have the unexercised option of changing the laws and the past.* This is not because of the strength of the laws or the past. It is because of their precision, and our relative imperfection. At best, our coarse-grained acts imply the falsehood of the laws or the past, but we are not actually in a position to falsify the laws or the past directly. (emphasis added)

This is a really enlightening passage on the nature of options in Maier's thinking and there are few things to say about it.

It is clear that in Maier's thinking options are, as he claims, coarse-grained, general kinds of powers (using the language of powers more metaphorically than literally). A coarse-grained power (or ability) would be one of a more general nature than a fine-grained power (or ability). For example, my ability to water my garden is *more general* than my ability to water my garden in a particular manner (e.g., quickly, with my left hand holding the hose, while whistling show-tunes). We might also carve up a particular action and suggest that when I water my garden, I actually exercise multiple fine-grained abilities rather than one general coarse-grained ability.

So, to use his example, if I draw a card, then I exercise an option (i.e., the option of drawing a card). But, the fact that I have an option to draw a card and, in actuality, the card happens to be a spade does not entail that I have an *option of drawing a spade*. The fact that I drew a spade also does not entail that I exercised *two* distinct options; that of drawing a card *and* drawing a spade, where drawing a spade (or exercising two options in one action) would be a more fine-grained descriptions of an option.

Let us dwell on this more. If I draw a spade, why should we not infer that I exercised *two* options (i.e., the option of drawing a card *and* the option of drawing a spade)? Well, the reason that Maier takes 'drawing a card' to be a genuine exercised option, but not "drawing a spade" (in a situation where he drew a spade) is because of *control*. Recall, On Maier's view, we must have a good amount of *direct control* over the outcome of our action (e.g., drawing a card or drawing a spade) in order to perform an action in order to have an option.⁸⁰ In the quotation above, Maier is claiming

⁸⁰ I thank John Maier himself for pointing this out to me and clarifying this reasoning through personal correspondence.

that we have enough control (required for action) in the case of drawing a card, but not enough control to specifically draw a specific card like a spade or a queen. To use the heuristic of probabilities again, if I form an intention to draw a card right now, all things being equal, there is a high probability I will draw a card right now. However, regardless of my intention, the odds I draw a spade is only .25. In this sense, I do not have the same kind of control over drawing a spade as I do drawing a card.

Maier also compares these two actions (i.e., drawing a card and drawing a spade) with options in a deterministic universe. By comparison, Option-Inheritance implies that, in deterministic universes, agents have the option to falsify the conjunction of the laws and state of the remote past *in virtue of* their option to raise their hand. And, an agent intuitively has control over whether or not she raises her hand. So the following kinds of control are supposed to be analogous or similar:

- (1) An agent's control over whether or not she draws a card (generally) is relevantly similar to an agent's control as to whether or not she raises her hand in a deterministic universe.
- (2) An agent's control over whether or not she draws a spade is relevantly similar to an agent's control as to whether or not she falsifies a proposition expressing the truth about some state in the past in conjunction with the laws of nature.

The idea that agents would have an option to falsify the conjunction of the laws and state of the remote past is wrong according to Maier because agents only have direct control to raise their hand (just like they do in drawing a card), but agents do not have direct control with relation to the *fine-grained implications* of raising their hand, which would be the falsification of the laws and the past (which would be analogous to our control of drawing a spade).

Rejecting the consequence argument is important for Maier's view of options. The upshot for Maier is that if he has reason to reject Option-Inheritance, he can reject the consequence argument. The consequence argument would have implied that agents in deterministic universes *only ever have one option* in deterministic universes. He does not think this is a convincing argument because (I) the Performance and Possibilities Principles do not entail Option-Inheritance (which he takes to be the key premise of the consequence argument) and (II) because Maier thinks it is wrong to suggest we would have an option to falsify the laws and the past because of the lack of control we have in that case relative to our control over coarse grained actions right in front of us.⁸¹

5.1.6 My Response to Maier

Without the consequence argument, Maier assumes (because it appears true regardless of determinism) that agents have multiple options in any given moment: some exercised options and some unexercised options. In a choice situation where one is deciding whether to draw a card or ask his neighbor for a card (playing go-fish), on Maier's view, if one draws a card, then one has the *unexercised* option to, say, not draw a card (even in a deterministic universe). But, one does not have the option to falsify the laws or past (for the same reason one does not have the option to draw a spade, no one has enough control over those states).

⁸¹ A related objection comes from Christian List (2019), who argues the chief problem with the consequence argument is that the modality relevant for determinism and the modality related for agency are (i) different and (ii) not combinable. So, on his view it would be simply a kind of modal category mistake to suggest something like "no one as the power to change the laws of nature and past." Maier does not suggest, though, that the agentive modality is not combinable in the way List suggests. I am assuming one craft these semantics into a broader modal semantics, which would be a good further project to undertake.

I am going to argue that Maier's second response to Option-Inheritance is problematic. Imagine I am playing go-fish and am deciding whether to draw a card or ask a neighbor if they have a card, when I decide to simply draw a card. My drawing a card is not counterfactually sufficient for my drawing a spade. There is enough chance built into card-drawing that makes that fine-grained aspect not in my control (even if I was trying to hoping to draw a spade). I agree with Maier this far; I can make sense of *this* claim that Maier makes (that I do not have control over the fact I drew a spade). I disagree, though, that "drawing a spade" is relevantly analogous to my control in falsifying the past and the laws (by raising my hand in a deterministic universe).

To modify an example from Chapter Two, imagine that I am given a vial of potion that will heal my disease even if the label reads: "Be aware: one out of a million takers of this medicine turn invisible." I think that it is reasonable to suggest that I do not here have the option of turning invisible, because my action is not actually sufficient for my turning invisible based on these odds. Though, I do have the option *to drink the vial*. However, if the label reads "Be aware: 60% of takers of this medicine turn invisible" then it seems intuitive to me that I also have the option of turning invisible (or at least the option of *trying* to turn invisible). Now, perhaps 60% is not high enough for control. But, what if the vial says "99% of takers of this medicine turn invisible"? In this case, on any virtually any criteria of control, I would have the option to turn invisible.⁸²

⁸² Although you do not have to analyze control through counterfactual sufficiency, I assume other methodologies to evaluate control would come out the same.

In a situation where I refrain from raising my hand in a deterministic universe, Maier wants to argue that I still have the (unexercised) option to raise my hand (even in determinism is true), even though I do not have the option of changing the past and the laws of nature. But, raising my hand *is counterfactually sufficient for the conjunction of the laws and past being false in a deterministic possible world*. The closest possible worlds to the actual world (assuming determinism) where I do an alternative action is one where the conjunction of the laws and past is false. Therefore, counterfactually, *falsifying the conjunction of the laws and past would be an option*. There would be a 100% chance I would change the laws or past of the universe if I had exercised that ability. Recall, Maier says, “At best, our coarse-grained acts imply the falsehood of the laws or the past, *but we are not actually in a position to falsify the laws or the past directly*” (2022, 145, emphasis added). In a deterministic universe, we would have an option to falsify the laws or past by virtue of the fact that we had the unexercised option to do otherwise (which, here, is to raise our hands). We might not be aware we have this option, but we are not always aware of every option we have in front of us.

I agree with Maier that there is something deeply implausible about having the option to falsify the past and the laws. This is why the FIXITY OF THE LAWS AND PAST principle in the P-MODAL CONSEQUENCE ARGUMENT is so intuitive. However, in a possible world where agents *have alternative options* and determinism is true, it would follow that agents have the option to falsify or change the conjunction of the laws and the past. The consequence argument-particularly the P-MODAL CONSEQUENCE ARGUMENT- is an argument showing that agents *cannot* have unexercised options in a deterministic world. Thus, that sort of counterfactual would have an impossible antecedent.

In summary, Maier's believes there are only two rules governing options: the Performance and Possibility principles because he believes the key premise in the consequence argument (i.e., Option-Inheritance) is false. He believes this premise is false because he does not believe agents in a deterministic universe with alternative (unexercised) options have the option of falsifying the conjunction of the past and the laws. I showed in my response that if we apply the same criteria for options that we do for action types like hand raising to falsifying the conjunction of the past and the laws, then we discover that, on Maier's account, we would have the option of falsifying this key conjunction. Maier is mistaken that we can call raising our hand an option, but not falsifying the conjunction of the past and laws an option (in this counterfactual situation) because it is importantly not analogous to the relation of having an option to draw a card, but not the option of drawing a spade.

Once we accept the soundness of Modal Consequence Argument (particularly the P-MODAL CONSEQUENCE ARGUMENT, but the B- MODAL CONSEQUENCE ARGUMENT is also sufficient), we learn that agents in deterministic or physicalistic universes do not have unexercised or alternative options. This would be a reason to adopt the following rule, which Maier briefly considers (2022, p. 138),

Historical Possibility Principle: If $S \text{ OPT } a$, then there is a possible world with the same laws and history as the actual world at which S performs an act of type a .

The Historic Possibility Principle is expressing the following truth: In a situation where I am deciding to raise my hand or refrain from raising my hand, assuming I have the option to decide either of these outcomes, but *actually* refrain from raising my hand, then there is a possible world

with the same laws and history as the actual world, except where I choose to raise my hand.⁸³

However, the Historic Possibility Principle would be gleaned as a result of the consequence argument, not an assumption in the consequence argument.⁸⁴

5.1.7 Are Options Primitive?

At this point, it seems to me as though the next natural question is: What are options metaphysically? Can we say anything more about them? Or, are they, as Maier argues, just metaphysically primitive notions. I am going to consider the possibility of placing Maier's view of options and abilities into a larger metaphysical framework. I will consider two possible such metaphysical frameworks: The first is from Helen Steward, who argues that ability claims are claims about distinct powers (one-way and two-way powers in particular). The second is from Kadri Vihvelin, who proposes a view of abilities as dispositions.

On Maier's account, we understand common ability ascriptions (e.g., Ben can play hockey) in terms of modal ability ascriptions which are dependent on the notion of options. Options are then course-grained outcomes that we decide on. For example, in a choice situation where I am choosing a flavor of ice cream, I have the *option* of choosing vanilla or strawberry. Next, I am

⁸³ If we do accept the Historic Possibility Principle, there is an interesting question about how we should do counterfactual analysis. This would be related to handling indeterminism and counterfactuals. For example, one could take some kind counterfactual skepticism (Hájek, 2021), or counterfactual indeterminacy (Kocurek, 2022) or a more traditional kind of counterfactual analysis (Stefansson, 2018).

⁸⁴ Sartorio (Sartorio & Kane, 2022) wants to argue that the proponent of the consequence argument assumes this sort of principle, I disagree with this assessment obviously. All the consequence argument proponent needs to assume is the notion that alternative possibilities (options) are required for free will and that one does not have the option to change the laws or past (or in the case of the B-Modal Consequence Argument, one needs to argue via Nomological Closure that we cannot change the microphysical states on which our mental states depend).

going to show how Helen Steward's theory of powers can be potentially combined with Maier's notion of options, although there is some cost with this combination.

According to Steward, abilities are fundamentally *powers*. Powers are like dispositions (i.e., a disposition is like a glass' fragility, a characteristic that can manifest under certain conditions) in that they are intrinsic modal properties of agents (and objects more generally). Steward categorizes these powers into two fundamentally different kinds. The first kind of powers are "two-way powers" (2012; 2020), which "are powers which an agent can either exercise or not at a given moment, even holding all prior conditions at that moment fixed" (Steward, 2020, 345). Two-way powers are unique kinds of abilities in that, when one exercises them, they make a difference to, or, in her words, "settle," the way history is going (Steward, 2013; 2020). When one suggests that an action or decision is "up to" the agent, Steward would suggest that what makes the said action "up to" the agent is that she can settle the way the power is exercised at that moment (Steward, 2020, 343).

One possible confusion here – which is brought up by Frost (2019) – is that the talk about ways in which a two-power can be exercised (namely, the *two* ways) can be misleading. It is not as if the title "two-way" should be understood as denoting as power manifesting *in two* ways. That would be problematic. Steward points out that "I might cry on Monday, Tuesday, etc.—but that doesn't make my power to cry a seven-way power" (Steward, 2020, 343). Two-way powers may be manifested *or not* given all the antecedent conditions of the exercising of that power (Steward, 2020, 348). Steward explains that "[t]he key is to think of the two 'ways' not as fundamentally different ways in which the (single) power might be manifested, but rather as ways things might proceed, given a situation in which an agent with a relevant two-way power is confronted with the opportunity to exercise it" (Steward, 2020, 354). She also explains that

What is special about it is that the number two is connected with negation. If I can $\emptyset$ and also not $\emptyset$, then that is two things I can do—exercise my agentive power, and also not do so. When I don’t exercise a given agentive power—e.g., the power to sing—that is not, of course, itself an exercise of the power to sing. Nevertheless, the possibility of this nonexercise is key to what makes an actual exercise of the power to sing attributable to me as its source. It is only because I need not sing, even when I have the ability and the opportunity to do so, that the singing is a true action of mine. (Steward, 2020, 349).

So, two-way powers are not powers that can be directed or manifested in two ways, but rather agential powers that *can* be exercised but not do have to.

Two-way powers can be understood further by contrasting them with Steward’s understanding of “one-way powers.” One-way powers “manifest whenever their manifestation conditions are realized, as a fragile glass breaks whenever it is struck” (Maier, 2020, S 5.2). One-way powers thus have sufficient and necessary manifestation conditions. Steward gives the example of “water [which] has the power to dissolve copper sulphate under certain conditions. But when those conditions are realized, water can’t do anything but dissolve copper sulphate. There is no alternative” (Steward, 2020, 345). When someone pokes me with a sharp object, I naturally give a pain-response to that poking (e.g., an “ouch”) because I am reacting to the stimuli. This is an example of a one-way power.

On Steward’s view, having two-way powers is essential for agency in general. Only creatures with two-way powers can perform agential actions and potentially have free will. If one is inclined toward Steward’s powers-metaphysic, one could read Maier’s notion of options as a way of

expressing two-way powers.⁸⁵ Maier (2022, pg., 136), himself, does seem to consider Steward's view, but he ultimately rejects it on the basis that it would entail agents in a deterministic universe do not really perform agential actions or have options at all. If agents cannot do otherwise (have no alternative options) in deterministic or physicalist universes, they do not have two-way powers (options) on Steward's view, and, therefore, are not agents at all. If we just imagine any agent, like Emily raising her hand in a faculty meeting, and assume for the sake of argument that this agent is in an indeterministic world, then it is puzzling why simply changing the laws would fundamentally change the agent (i.e., Emily) into a non-agent. This is an interesting possible interpretation of options, however it is not necessary.

In summary, thinking of options as a kind of possible exercising of two-way powers is a possible addendum or move one could make. The benefit of this combination is that powers are arguably metaphysically clearer than options, so options can more easily fit into one's ontology. However, the tradeoff is that the Modal Consequence Argument (of either B or P types) would seem to entail that agents in physicalistic or deterministic possible worlds *would actually be non-agents (i.e., objects)*. Incorporating OPTIONS-BASED ALTERNATIVE-POSSIBILITIES into the modal consequence

⁸⁵ Although, perhaps, this would make Maier's notion of options reductive, but they would just reduce to two-way powers would not readily reduce to powers as these are very different than powers generally. So, options would reduce one more level and then stop. There is nothing obviously contradictory about viewing two-way powers in a possible world evaluation.

argument (of either B or P types) by itself does not have this specific entailment, which is arguably the more intuitive inference.⁸⁶

We can briefly contrast Steward with another view of abilities: the dispositional view advocated by Kadri Vihvelin (2013). Vihvelin's view is a dispositional view. For example, my coffee mug is breakable. It is liable to break under certain foreseeable circumstances, and we often take this breakableness to be a kind of intrinsic property. This is relevant because we typically think dispositions are there even if they are not manifesting in front of our eyes. As Vihvelin (2013, p. 171) says,

In relying on this [dispositional] knowledge, we assume that dispositions are relatively stable characteristics of things that, *ceteris paribus*, continue to exist when they are not manifested. We also believe that there is a sense in which something with the disposition to X can X even during times that it is not Xing, and even if it never X's.

For example, my coffee cup is breakable (because it is fragile) even if it is in bubble wrap (so it will not break given its conditions). On Vihvelin's view, persons have dispositions as well. "Some people are color-blind; others have normal color-vision; a lucky few can discriminate shades and hues the rest of us cannot see. Some people can understand French; others can't. Some people are easygoing, others are more easily provoked, and some are hot-tempered. And so on" (Vihvelin, 2013, p. 172). According to Vihvelin, abilities are dispositions, as she says, "[t]o have an ability

⁸⁶ My argument shows that these agents are not *free*, not that they are non-agents.

to act is to have a disposition or bundle of dispositions” (Vihvelin, 2013, p. 172). She further notes (Vihvelin, 2013, p. 172),

There are some striking similarities between our abilities to act and the dispositions that we share with natural objects. We have empirical knowledge of both. Abilities, like dispositions, don’t usually pop into existence only on the occasion of their exercise or manifestation. Nor do they usually go out of existence simply because a person is not exercising them. A person may lose her ability to speak French after not speaking French for many years, and she may lose her ability to walk in a straight line when she is very drunk, but she does not lose these abilities every time that she ceases to exercise them. Finally, we believe that there is a sense in which a person with the ability to do X can do X even when she’s not doing X, and even if she never does X. The monk who has taken a vow of silence remains able to speak; he can speak, even though he never does.

If we take this view and try to see common ground for Maier’s semantics, we could interpret options as something like the possible manifestations of dispositions. However, there is an obvious problem, which Maier himself points out, which is that the consequence argument could still threaten one’s *options* on this view, but not one’s *abilities* (Maier 2015, p. 133). So, whatever benefit there is to this broader metaphysical framework, the benefits of Maier’s agentic modalities would largely be lost (e.g., the Robustness Objection). And, options seem to be more important than abilities. So, Vihvelin’s dispositional view of abilities is not a great fit for Maier’s agentic modalities.

In this particular section, I showed how one could theoretically fit Maier’s notion of options within Steward’s larger metaphysical project, however this is not a necessary move one has to make in accepting Maier’s agentic modalities. From a bigger picture, this section has shown that by

incorporating OPTIONS-BASED ALTERNATIVE-POSSIBILITIES as an assumption in the B-MODAL CONSEQUENCE ARGUMENT there is a very robust explanation as to why we should accept that alternative possibilities matter for free will. In the next section, I will consider a further objection though.

5.2 The Frankfurt-styled Objection to The Consequence Argument

I started this chapter by discussing premise (4) of the B-MODAL CONSEQUENCE ARGUMENT. Recall, Premise (4) is

Premise (4): Necessarily, if there is nothing that Emily (or anyone) can do that would make the state of affairs of Emily deciding to raise her hand at t not obtain, then Emily cannot freely choose to raise her hand at t .

Given the discussion in the last section, we can now see that one way to infer Premise (4) is from the following principle:

OPTIONS-BASED ALTERNATIVE-POSSIBILITIES: A creature C is free in ϕ -ing *only if* C , when choosing to ϕ , had the option to decide to $\neg \phi$.

In order for Emily to freely choose to raise her hand at that faculty meeting she must have the option to refrain from raising her hand, and having the option to refrain from raising her hand (at t). If she does not have the option, then she cannot do anything to make that state of affairs (of Emily deciding to raise her hand at t) not obtain.

One obvious objection to the B-MODAL CONSEQUENCE ARGUMENT (or the P version) is to reject premise 4 entirely due to the success of so-called Frankfurt cases. Frankfurt cases are principally inspired by Harry Frankfurt's article (1969) where he tried to show that there are instances of

agents freely deciding or acting *without having alternative possibilities*.⁸⁷ Frankfurt cases are relevant for all varieties of the consequence argument (including the B-MODAL CONSEQUENCE ARGUMENT) because they provide a potential counterexample to a key premise in the consequence argument, which in my version is premise (4). As Al Mele puts this objection, “If some Frankfurt-style story works, it would show that an action doesn’t need to [have an alternative possibility] to perform it freely, in which case the Consequence Argument would bite the dust” (Mele, 2022, p. 40). The consequence argument would be threatened, not because the Beta Rule or Fixity Principles are false, but because the premise that the present is fixed (Premise 3) is shown not to entail that free will is impossible (Premise 5). In other words, the consequence argument (of any variety) *can entail* premise (3), but the argument does not entail that Emily does not have free will (Premise 5).

The plan for this chapter is fairly straightforward. I am going to consider four different versions of Frankfurt cases (and their corresponding arguments, which all have a similar structure). In §5.2.1, I am going to consider a very basic ordinary Frankfurt case in order to show the structure of the

⁸⁷ Frankfurt cases are also commonly used to show that agents can be *morally responsible* without alternative responsibilities. That variation of the argument is relevant, but is notably different than what I am aiming at because moral responsibility is different (although related) to free will. I am going to focus on free will for clarity and simplicity, but many of the same discussion points in this section can be applied to the discussion of Frankfurt cases and moral responsibility. Frankfurt’s article is famous in the free will literature, as Van Inwagen indicates: “The classical era may usefully be regarded as falling into two parts – before and after the publication of [Frankfurt’s] essay” (2017, p. 215). As Vihvelin (2013, p. 94) points out, in terms of impact and scope, Frankfurt cases are just as impactful in free will as Gettier cases are in Epistemology:

It’s difficult to explain, to someone not working in this area, just how peculiar the situation is. On the one hand, Frankfurt stories, as they have come to be called, have had an impact in free will circles that is comparable to the impact of Gettier stories in epistemology. On the other hand, after over forty years of debate, it is still controversial whether Frankfurt or any of his followers have succeeded in providing a genuine counterexample to PAP.

argument, and then, argue that there are a few different possible responses to such a case. Following Steward (2013) and Alvarez (2009), I will argue that probably the best response is to assume that some of our alternative actions are moderately fine-grained (§5.2.2). Then, I will move to a more standardized version of the Frankfurt case, which I take from John Martin Fischer (1982). In §5.2.4, I explain what I take to be an effective criticism of this sort of case originally from Robert Kane (1985), known as the Dilemma Defense. The Dilemma Defense argues that Frankfurt cases must either be assuming determinism or indeterminism, but if one assumes determinism one is question-begging against the consequence argument and if one assumes indeterminism, the case is not successful. Al Mele and David Robb take the second horn of that Dilemma to task and offer a version of the Frankfurt case that assumes indeterminism and is allegedly successful. I discuss this case in §5.2.5, and then offer two responses in §5.2.6. Finally, I deal with a contemporary version of the argument from John Martin Fischer (2022), who argues that the addition of God can make an indeterministic Frankfurt story successful (§5.2.7). I consider this argument and offer a response in §5.2.8 showing that it is, in fact, still not successful.

Having shown that no Frankfurt-styled story is successful without question-begging, I will have shown that a Frankfurt-styled Objection is not successful to the P or B MODAL CONSEQUENCE ARGUMENT.

5.2.1 The Ordinary Frankfurt-styled Argument

There are several paradigmatic kinds of Frankfurt cases. All alleged cases are supposed to be “circumstances that constitute sufficient conditions for a certain action to be performed by someone and that therefore make it impossible for the person to do otherwise, but that do not actually impel the person to act or in any way produce his action” (Frankfurt, 1969, p. 830). By contrast, a kind of case where a mad-scientist directly manipulates a person via mind-control would

be a way that “impel[s] the person to act.” I am going to consider one of these kinds of cases here, then later in the chapter, I am going to consider the other varieties of the Frankfurt case (and this *other* varieties are probably more problematic).

The first kind of paradigmatic Frankfurt case would be like the following:

Drowning: Ben is a generally moral person who is walking along the beach one morning enjoying the sunrise when he notices his child drowning in the ocean a short distance away. Instantly, Ben decides to leap into the water to save his child.

This kind of case is simple and intuitive.⁸⁸ It is a case where a person has clear and overwhelming reasons for some decision or action. When persons have clear and overwhelming reasons for some decision or action, then that decision or action seems necessary. Assuming Ben is an ordinary moral person, it seems impossible for Ben to decide to do anything else in this scenario than to save his child. Of course, we can imagine Ben refraining from deciding to save his child, but Ben would be a moral monster in such imaginative alternative cases, a completely different person.⁸⁹ If we then take this case at face value, Ben seems to have overwhelming reasons to perform the action of saving his child and no clear alternative possibility. Thus, we can ask whether or not Ben has any real options in this case? In other words, is Ben deciding to exercise one option among multiple

⁸⁸ This case resembles Steward’s burning building case (2013, 183-184) or Dennett’s Luther case (Dennett, 2015, p. 144).

⁸⁹ It is possible to imagine Ben refraining from deciding to save his child in scenarios where there is some greater good associated with his refraining, and in those cases, Ben is not a moral monster. For example, suppose when Ben sees his child drowning, he *also* sees a nuclear bomb device next to him ticking down from 10 (to 0), where all Ben has to do is type in the code to stop the explosion (which is also present).

options in this case (because it does not seem like Ben *cannot* decide this particular action)? There are three responses I am going to discuss here.

The first possible response to such a case is to suggest that Ben does not actually exercise his free will in the Drowning case. Why would one suggest this? Well, some free will theorists suggest that free will is *only* exercised on limited occasions. Richard Swinburne (2013, p. 197) argues that free will really only manifests in situation where “our strongest moral beliefs conflict with our strongest desires.” It is in these torn-situations, we exercise our free will.⁹⁰ And, this is not a torn-situation. Cases such as Drowning, where persons have very strong moral reasons and natural desires for same attainable outcome, would be cases, according to this view, where persons (more-or-less) *react* to the situation. If you poke me with a sharp object, I will naturally react by a pain-reaction (e.g., saying “ouch”). Ben’s decision in Drowning would be analogous to my pain-reaction in a certain respect. However, this type of free will theorist is not suggesting that Ben’s decision is *just* like my pain-reaction, it still is a kind of action.

Some readers may think this response is counterintuitive because Ben is obviously morally responsible for his decision and action in Drowning. Recall, though, that OPTION-BASED ALTERNATIVE POSSIBILITIES is a principle about free will, not Moral Responsibility. One could suggest that Ben (and persons like Ben) could be morally responsible for his action because there was some prior freely willed decision (involving alternative possibilities) in the causal history

⁹⁰ One might wonder how often these particular situations occur on this sort of view. To this question, Swinburne suggests the average human only has one such occasion a day, but, he notes, “there are conscientious people who are faced with such choices much more often than that, and other people less sensitive to moral dilemmas who face difficult moral decisions only perhaps once a month” (ibid).

leading up to his decision to save his child. For example, Van Inwagen (2015, 23) suggests the following principle

[Moral] Principle of Alternative Possibilities: “A person is morally responsible for some of the consequences of some of his acts only if that person is sometimes (i.e., at certain points in his life) able to do otherwise.”

Given [Moral] Principle of Alternative Possibilities, Van Inwagen claims that is not even possible to use the “general idea” of a Frankfurt case to construct a successful counterexample (2015, pp. 23-24). One will realize that, given [Moral] Principle of Alternative Possibilities, Ben (in Drowning) could be morally responsible even if there is no present alternative responsibility, and even if Ben did not *freely* save his child.

This is a perfectly legitimate and effective response. However, the tradeoff is that it is quite intuitive to think that Ben does *freely decide* to save his child. Taking this approach would effectively cut off many of the intuitively paradigmatic cases from which we conceptually analysis free will. There is a seemingly arbitrary line between cases where we exercise our free will and cases where we do not.

A second possible response is from Christian List (2019, p. 24-25) who argues that the proper interpretation of these Frankfurt-styled cases (like drowning) is *not* “that it was literally impossible from him to do otherwise. Rather, he was not able to do otherwise *without sacrificing his integrity*. Had he done otherwise, which was within his abilities, he would have betrayed his values and probably regretted his choice” (2019, p. 24). So, on List’s account, we should not grant, *prima*

facie that Ben (in Drowning) has no alternatives. Ben *could be* a moral monster and he is choosing not to be one.⁹¹

This is, again, an effective response in that it does answer the relevant question, but more needs to be said here as to how we know that Ben really has this alternative option. Can we really not get to the moral point in life where voluntarily allowing a child to die (for convenience) is simply not an option? I suppose one's moral theory could allow this, but I am going to discuss a different sort of response next that, I think, is more intuitive.

5.2.2 Helen Steward and Maria Alvarez's Response

The third sort of response comes from Steward (2013, p. 184-185) and Maria Alvarez (2009, p.73) directly, but we can use the same line of thought from the perspective of our Maier-inspired view of OPTIONS-BASED ALTERNATIVE-POSSIBILITIES, who both argue that in Drowning Ben (the protagonist) should have exercised a two-way power when he decided to try to save his child. In Maier's view, we could say that Ben exercised *an option in saving the child*. Therefore, the obvious question is how this would be possible given our ordinary intuition that Ben could not have done otherwise. Steward and Alvarez' solution is to distinguish more general (i.e., coarse-grained) descriptions from more specific (i.e., fine-grained) descriptions of the protagonist's action or decision. And, when we specify the more fine-grained descriptions (although it does not have to

⁹¹ In Maier's notion of options, too, we are not always aware of all of our options, so it could be that we just never think about the alternative of letting our child die in such a circumstance, but it is a real option.

the most fine-grained descriptions possible, it can be *moderately* fine-grained), we should be able to spot the (implicit) relevant alternative possibilities.

Steward (2013, p. 185) argues that agency requires that we have at least some of the powers to bring about the fine-grained element- the processes, organization, structure of our action. To utilize her example, perhaps I want to wash my hands in a situation where my hands lack feeling because they got too cold. There are certain things I could not do here, I could not feel the temperature of the water as it came down (assuming my hands were *that* numb), but I could still rotate my hands in certain directions, speed up the motion and so forth. This is different than a situation where none of the fine-grained elements (the undergirding processes and organization of my action) are in my control. Steward gives the example of being a compulsive washer; she notes that if she were a genuine compulsive washer it is conceivable she could *not* rotate her hands differently, change the temperature of the water, stop herself from washing (e.g., “if set upon by a vicious dog”), or do any fine-grained element any different (ibid). She concludes, “If I simply possessed none of these powers to settle the particular details of the action I execute, I simply could not be said to be its agent at all” (ibid). On her view the compulsive hand-washer would be a kind of machine or automata, rather than a genuine agent. Therefore, on her view, one needs alternative possibilities for agency itself. If one has the agential ability to ϕ , then, according to Steward’s view, one must have the power with respect to at least some of the particular undergirding processes and organization of ϕ -ing (e.g., moving one’s fingers in order to pull a trigger).

Applying this to Ben’s circumstance in Drowning, Ben intuitively could not decide to watch his child die as Ben is not a moral monster. However Ben can decide, for example, *how* to save his child. He has to decide whether to, for instance, kick his shoes off immediately or dive into the water with his shoes. He has to decide the best course of action to save his child. In his decision to

save his child, there are many specific details about the process that would still have open possibilities even if the broader decision (i.e., to save his child) is necessitated (because of Ben's moral character and the circumstances). In essence, Steward is envisioning the ways in which two-way powers are manifested should be more fine-grained, or specific. And, in Maier's option language, we can say that there are genuine (moderately) fine-grained options available for Ben here. More formally, on her view, if an agent performs some action ϕ at t , then her ϕ -ing must have some relevant level of description such that the agent could have not ϕ 'd at t . By "some relevant level of description," she means some aspect of the action or decision that is normally subsumed by the general description (2013, p. 185).

With this more fine-grained description of the relevant options, cases like Drowning are not going to be cases where the agent cannot do otherwise. These would rather be cases where agents do not have options at the most general level of coarse-grained descriptions. But, this is not surprising. Ordinary persons are not superheroes or omnipotent beings so there are always natural limitations. In some circumstances (like Drowning), there are further limitations on the options available to the person.

This kind of case is a really helpful case to start with. It is simple and clear. However, the next few examples of Frankfurt-styled cases are less ordinary, but just as important to discuss.

5.2.3 The Standard Frankfurt-styled Argument

To begin, I am going to present the standard elements of a Frankfurt case, and then introduce a common version from Fischer (1982). After this, I will present how we could turn this case into an argument against the premise (roughly) that, necessarily, if Emily does not have alternative

possibilities, then she cannot have free will (Premise 4 of the B MODAL CONSEQUENCE ARGUMENT).

There are three critical aspects of a Frankfurt case:

- (I) Firstly, there is some protagonist (typically named Jones) who is deciding between options (i.e., x or z), and then decides on one option (e.g., x). One caveat to these options and the protagonist's decision is that they have to be such that if the reader just took this the decision and options in isolation, the intuitive judgement would be that the protagonist is *freely deciding* x (so there can be no manipulation or coercion present). All Frankfurt cases and variations have this first element.
- (II) Second, unbeknown to the main character, there is an antagonist character (often a scientist named Black) who wants one particular outcome (e.g., x), and it is the same outcome that the protagonist ends up choosing on his own. However, the antagonist, prior to the outcome, is not sure if the main character will actually choose the option he wants, so, as a result, plans to *manipulate (extremely)* the main character, if the main character chooses the outcome the antagonist does not wish to occur.
- (III) Thirdly, the antagonist has the ability to both *predict* and *manipulate* the protagonist by some manner. These functions are typically explained by some causal mechanism (e.g., a neurological device), but could theoretically be explained by some sort of miraculous power (e.g., occult or miraculous powers). This third aspect of Frankfurt-cases is the aspect that is quite different between prospective imaginative cases (i.e., some are overdeterminative some are preemptive).

All three aspects need to be present in any Frankfurt-styled case. There are dozens of proposed examples of Frankfurt cases in the literature.

I am going to show one case from John Martin Fischer, originally written in (1982), that is a clearer variation of Frankfurt's original case (1969):

Fischer's Frankfurt-Case: Black is a stalwart defender of the democratic party. He has secretly inserted a chip in Jones's brain that enables Black to monitor and control Jones's activities. Black can exercise this control through a sophisticated computer that he has programmed so that, among other things, it monitors Jones's voting behavior. If Jones were to show any inclination to choose to vote for anyone other than the democrat, then the computer, through the chip in Jones's brain, would intervene to assure that he actually chooses to vote for the democrat and does so vote. But if Jones chooses on his own to vote for the democrat (as Black, the old progressive would prefer), and he does actually so vote, the computer does nothing but continue to monitor – without affecting – the goings-on in Jones's head. Now suppose that Jones chooses to vote for the democrat on his own, just as he would have if Black had not inserted the chip in his head. It seems, upon first thinking about this case, that Jones can be held morally responsible for his choice and act of voting for the democrat, although he could not have chosen otherwise and he could not have done otherwise. (Fischer, 2022, pp. 1186-1187)

In this case, we can clearly see all the three elements mentioned above. First, Jones has an important decision in front of him, and he chooses to vote for the democratic candidate on his own. Second, Black wants Jones to vote democrat and is trying to predict what he will do. Thirdly, Black is planning on manipulating Jones (via this chip and computer) if Jones does not vote democrat.

If we take Fischer's Frankfurt Case and try to turn it into an argument against Premise (4), we need to begin by taking into account that, if Black were not present in any sense, then the intuitive inference would be that Jones freely decides to vote for the democrat. Let us refer to the world in

which Jones votes democrat, the actual world of the scenario as Fischer specifies, as w , and a world exactly like w except where Black is totally absent as w^* . Thus,

P1: Jones' choice to vote democrat would have been free in a possible world w^* identical to the one mentioned except where Black was not present. (Assumption)

However, in w (the actual world of the scenario), Black does *not* actually do anything to Jones seemingly relevant for his decision. Jones' decision seems like it is the same (i.e., his decision goes through the same processes) in both w and w^* . Thus, we can infer that Jones' decision (in w) is not fundamentally different than it would have been if Black had been absent:

P2: There is no relevant difference between Jones' decision to vote democrat in w and w^* . (Assumption)

P3: Therefore, Jones' choice to vote democrat is free in w . (From 1, 2)

However, it is also asserted in this situation (w) that, because of Black, in w , Jones "could not have chosen otherwise and he could not have done otherwise" (2022, p. 1187). Thus, if we read that in the same manner, we read the N-Operator relevant for the B-MODAL CONSEQUENCE ARGUMENT, we can infer that:

P4: There is nothing Jones can do in w that would make Jones' deciding to vote democrat not obtain. (Assumption)

So, if we take P3 and P4 together we get,

P5: There is nothing Jones can do that would make Jones' deciding to vote democrat not obtain and yet Jones' decision to vote democrat is free in w .

If P5 is true and given that it conflicts with Premise (4) in the B- MODAL CONSEQUENCE ARGUMENT, we seemingly can then infer: Premise (4) is false.

Note, that the scenario in question is supposed to imply two separate truths. First, given that is no difference between Jones' decision in w and w^* (that is, between a world where Black is present and one where Black is totally absent), Jones' decision in the actual sequence is supposed to be free. Second, in w , Jones is supposed to be unable to do anything other than what he actually does because, in the alternative sequence, Black would have manipulated Jones. This is due to, in Fischer's case, Black's neurological chip and some computer program.

This is a standard Frankfurt case and a way the case can be formally turned into an argument against Premise (4). In the next section, I will discuss why I believe this argument fails.

5.2.4 The Dilemma Defense

In this section, I will explain a common response to the Frankfurt-styled argument (as presented in 5.2.1) called the Dilemma Defense. I will argue that this defense is successful in showing that either P3 (that Jones is free) or P4 (that Jones could not do otherwise) must be false in the Frankfurt-styled argument.

The objection I will make to the Frankfurt-styled argument (as presented above) is a dilemma first presented by Robert Kane (1985). The dilemma starts with a question: Is determinism true in w (recall, w is the world like w^* but where the antagonist is actually present), or is indeterminism true in this scenario? These are two different horns. Obviously, either determinism or indeterminism is true in w . So, which is supposed to be true?

On this first horn, if determinism is true, then the intuitive judgment readers infer about Jones (i.e., that he is free) is threatened by the consequence argument. The judgment that Jones is free (P3) is only inferred from the pre-theoretic judgment that Jones would be free if Black were not present in that world (P1 and P2). However, if determinism *is true* in w , then there is a difference between w and w^* (because determinism is not factored into the pre-theoretic judgment of Jones' freedom in P1). It is question-begging to assume that this difference is not relevant. The proponent of the modal consequence argument would have a reason to think the difference is relevant (the P-MODAL CONSEQUENCE ARGUMENT) and, thus, P2 (there is no relevant difference between w and w^*) is false.

However, if we consider the other horn of the dilemma (i.e., indeterminism), it is unclear how the device would be able to predict Jones' decision, and, thus, genuinely take the alternative possibilities away from Jones (so, P4 is false). To see why, imagine there was a genuinely indeterministic coin. A computer could not predict what side the coin *would* land on? So, how could it predict Jones' decision? It does not seem to be able to do so. Let me spend a moment talking about each of these horns further.

Let me begin with the first horn. Robert Kane (1996, p. 142–145) has argued that the only way for Black's device to be able to detect Jones' inclinations is if we assume causal determinism holds true in the Frankfurt Case. As explained earlier (reference), determinism is the metaphysical view that a completed state of the universe at one time in conjunction with the laws of nature entails all other completed states of the universe at all other times. Stewart Goetz, building on Kane's original argument, explains,

[The Frankfurt-style example] creates the appearance that it is Black's device [in the alternative sequence of events] that makes it the case that Jones is not free to choose

otherwise. This appearance is *illusory* because without the obtaining of causal determinism in the actual sequence of events, the device cannot prevent Jones from making an alternative choice, and with causal determinism in the actual sequence of events it is not the device that prevents Jones from making an alternative choice. In short, if Jones is not free to choose otherwise, it is because of the occurrence of causal determinism in the actual sequence of events and not because of Black's device in the alternative sequence (2005, p. 85)

Goetz's point here is that, if determinism is true, the underlying reason why Black's neurological device works is that Jones' decision to vote for the democratic candidate is *predictable*, and it is predictable because it is *fully determined* (i.e., determinism). The Frankfurt-styled argument is predicated on P3 (that Jones is free in the actual sequence in *w*), which is inferred on the pre-theoretical intuition that Jones would be free in a similar circumstance absent Black (and his device). However, if Black can only preempt Jones' choice via a device that, in turn, can only work if determinism is true, then the case is question-begging. The only way to discern whether or not Jones could be free in the actual sequence is to first decide whether determinism and free will are a problematic combination (i.e., we must determine whether or not the consequence argument works).

What about the indeterministic horn of the dilemma? Well, on this horn there is a different kind of problem: Imagine that indeterminism is true in the case presented, but alternatively Jones *did try to decide to vote for the republican candidate*. This is the alternative sequence. Of course, the moment (T) Jones decides, or begins to decide, or try to decide to vote for the republican candidate, Black's device switch's on and manipulates him to vote for the democratic candidate. The reason the device does not activate before T (i.e., T-1) is that Jones decision is indeterministic, similar to

a genuinely indeterministic coin. Whatever neurological pattern is present at T-1 that probabilistically indicates what Jones' decision will be at T, it will not perfectly predict Jones' decision. Jones could have the neural pattern consistent with his voting democrat at T-1, and then still change his mind to vote republican. Thus, a counterfactual like the following will be true in the alternative sequence: "If Jones had begun deciding to vote for the republican, then Black would have manipulated him to vote for democratic candidate." But, this counterfactual seems to indicate that in the alternative sequence Jones would have had to begin to decide (or try to decide) in order for the device to switch on. As Van Inwagen (1978) pointed out, this is in of itself a kind of alternative possibility present in the alternative sequence.⁹² It is difficult to imagine a neurological device would be able to avoid this possibility given indeterminism. (This is what is often called the Flicker Strategy within the indeterministic horn.)⁹³

So far, I have argued that there are two horns every Frankfurt-styled Argument must reckon with. It seems clear that if the case assumes determinism, it is question-begging. However, I have also argued that if it assumes indeterminism, the case in question might not be conceivable. There two prominent cases that Frankfurt-proponents use to respond to this second horn. In the following sections, I will discuss both of these cases.

⁹² This is called the "Flicker of Freedom" strategy often. There are various contemporary versions. Philip Swenson and Justin Capes (2017) have pushed this defense in interesting new ways as well as fits with my moderately fine-grained response to the Ordinary Frankfurt Case.

⁹³ See Pereboom's *Tax Evasion* example for this kind of attempt (Pereboom and McKenna, 2012). These sorts of examples focus on moral responsibility more than free will, so I will side step these to a more directly obviously relevant Frankfurt Case in the next section.

5.2.5 Mele and Robb's Frankfurt Case

In response to the Dilemma Defense, the Frankfurt-proponent could propose a case showing the protagonist causing the relevant outcome in the choice situation where *indeterminism* is true, yet the protagonist is unable to do otherwise. Here is one such case (paraphrased for clarity) from Mele and Robb (1998, pp. 101-102) that is slightly more complicated, but does contain the elements just mentioned:

Bob: Bob lives in an indeterministic world. At t , the mischievous neurosurgeon Black is back (!) and wants to force Bob to steal Ann's car. Black has implanted a device in Bob's head (much like the Standard Frankfurt Case), and at t_1 (some time after t), Black activates the device, which starts a deterministic process P . P will ultimately culminate at a later time, t_2 , in forcing Bob to decide to steal Ann's car. P is a process such that (I) P will *deterministically* result in Bob's decision to steal Ann's car, (II) P is such that Bob is wholly unaware of P because Black has screened P from Bob's consciousness, and (III), P is a type of causal process that is not causally efficacious at t_2 (although it is before t_2) if Bob decides *on his own accord* to steal Ann's car at t_2 . In other words, P can be *overridden* by any indeterministic process of Bob's own will. To be clear, it is not that P is a potential preempting cause, but rather P is such that it will be effectively causally overridden by Bob's decision at t_2 (if Bob decides to steal Ann's car) in a way that is not jointly causal at all. (Black is unaware that Bob might actually decide on his own to steal Ann's car at t_2 , so this aspect of P does not concern him.) As it happens, what actually happens is that P is activated, but Bob decides on his own to steal the car "on the basis of his own indeterministic deliberation about whether to steal it, and his decision has no deterministic cause"

There are several elements to this story that are unique. *P* in this circumstance is particularly unique in that it is active in the actual sequence, but not the actual cause of Bob's stealing the car because Bob's indeterministic decision overrode it. Mele and Robb offer a couple of different analogies to make sense of this sort of causal overriding (see Mele and Robb, 1998, pp. 104). Probably the simplest analogy they offer is a seesaw (see, Mele and Robb, 2003).

In this analogous seesaw, imagine the center fulcrum is slightly off-center to the right and there is a cup on both sides of the seesaw. Now, the seesaw going down on either side is analogous to Bob's decision to steal Anne's car. Further, imagine a deterministic bb landing in the cup on the left at a moment *t* (this is analogous to *P*) and an indeterministic bb landing in the cup on the right at *t* (this is analogous to Bob's indeterministic decision to steal Anne's car), and in this event, the seesaw will obviously go down on the right side (because of the fulcrum being slightly off to the right). Although it is true that if the right bb had not been present, the seesaw would have still gone down (on the left side). There are two key elements of *P* in Mele and Robb's Frankfurt-case (i) *P* must be sufficient to make Bob steal Anne's car, but (ii) it must, in these specific circumstances at least, be causally differential to Bob's indeterministic decision if Bob chooses to steal the car on his own. The bb's in the seesaw case are like this. The left bb is sufficient to make one of the sides of the seesaw touch the ground, but, in these circumstances, is differential to the right bb (which is indeterministic).

We can write Mele and Robb's potential argument (just like in the Standard case) using this case in the following way:

P1: Bob's choice to steal Anne's car would have been free in a possible world *w** identical to the actual world *w* mentioned, except where Black was not present. (Assumption)

P2: There is no relevant difference between Bob's decision to steal Anne's car in w and w^* . (Assumption)

P3: Therefore, Bob's choice to steal Anne's car is free in w . (From 1, 2)

P4: There is nothing Bob can do in w that would make Bob's stealing Anne's car not obtain.
(Assumption)

So, if we take P3 and P4 together we get,

P5: There is nothing Bob can do in w that would make Bob's stealing Anne's car not obtain
and yet Bob's choice to steal Anne's car is free in w .

In this argument, the problem is going to be P2 or P4, which I will explain next.

5.2.6 Two Responses

There are two good reasons to reject P2 in this version of the argument. The first comes from David Palmer and Carl Ginet (2010) and the second is admitted by Robb himself as a problem (2020).

Briefly, before I mention the first problem. Recall the seesaw analogy: The bb's in the right cup are supposed to be analogous to Bob's indeterminate decision in some manner. Now, it is unclear the best way to think about these bb's being indeterministic in this analogous circumstance. Perhaps one way to think about it is simply imagining a person deciding to drop the bb's in the right cup, while there is deterministic process (e.g., a machine) dropping the bb in the left cup at the same time. However, it is not that the right side of the seesaw touching the ground is analogous to Bob's decision to steal the car, it is the more general event-*either side* of the seesaw touching

the ground—that is supposed to be analogous with Bob’s decision. Now, I can explain the first problem.

The first problem with this argument is that having the more general event be analogous to Bob’s stealing the car creates a dissimilarity, as Palmer and Ginet recognize (2010, p. 445),

But [the seesaw example] does not entail what Mele and Robb need. Recall that the event type of Bob’s deciding to steal the car is modeled, not by the more specific event type of the right end going down, but by the less specific event type of an end (one or the other) going down. And when an event of this less specific type results from simultaneous impacts, as it always must, it is deterministically caused by the event of those simultaneous impacts. That is, causal laws governing the mechanism dictate that, given an event of the type simultaneous impacts in the two cups, it will always cause an event of the type an end goes down.

Palmer and Ginet’s claim is that, by analogizing the more general event (i.e., either side of the seesaw hitting the ground) with the more specific event (i.e., Bob deciding to steal Anne’s car), we have effectively assumed determinism. Here, it does not matter that one *bb* has an indeterministic causal antecedent, given all the casual antecedents of the more general event, the seesaw hitting the ground was bound to happen.

But, there is a second objection that needs to be brought up: So, what would have happened if Bob chose to have refrained from stealing Anne’s car (perhaps he chose to propose to Anne!)? After all, if *P* can be overridden by an indeterministic process from Bob, if Bob had chosen otherwise, should the alternative sequence not have Bob doing something else? It seems so, since *P* causally

defers to Bob's own decision in the actual sequence. If this is so, then Bob has alternative options, and P4 is false.

Mele and Robb (Mele & Robb, 1998, pp. 104-105) anticipate this question, and note that it is natural to wonder what would happen at t_2 if P and Bob's indeterministic process (i.e., x) were to diverge at t_2 . Further the authors imagine our brains, hypothetically, as having "decisions nodes" such that if a deliberative or neural process (like x or P) were to "light up" such a node, the agent would perform that action. So in this circumstance, $N1$ represents the decision node of Bob's decision to steal the car and $N2$ represents refraining from stealing the car. They say that there are three notable possibilities. The first two are straightforward:

Situation 1: If P hits $N1$ at t_2 and x does not, then P would cause Bob's stealing the car

Situation 2: If P and x hit $N1$ at t_2 , then x would cause Bob's stealing the car

In Situation 1, P causes Bob's action. In situation 2, a kind of overdeterminism situation, x would cause Bob's stealing the car because P is deferential to x causally. The final relevant situation is the one we are interested in:

Situation 3: If P hits $N1$ at t_2 and x hits $N2$ at t_2 , then P would cause Bob's stealing the car.

In other words, in Situation 2, P is *causally* deferential to x - which is why x causes Bob's stealing in that situation- but in Situation 3, P preempts x . The authors note that of course, "readers would like a story about why it is that although x would preempt P in the former situation, P would prevail over x in the latter" (ibid, 105). Here is there explanation, P is a kind of process that, by t_2 , has "neutralized" $N2$ and "all of the nodes in Bob for decisions that are contrary to a decision at t_2 to steal Ann's car" (ibid). So, P is not just like pumping Bob with some kind of influencing desire to

steal Anne's car, it is something that actively changes Bob's brain such that (i) any alternative decision at t_2 is blocked, neutralized, or made temporarily impossible, and (ii) will be causally deferential to x , on the small chance Bob's indeterministic process x hits $N1$. This can be difficult to imagine, but I will assume it is possible for the moment.

So, in order to block the alternative sequence, Mele and Robb imagine that the causal work P is doing before t_2 is blocking or neutralizing all other possible decision nodes in Bob's brain in the actual sequence. Here, it is quite obvious that this addition to the Bob scenario will make Bob having alternative options impossible (and so $P4$ correct). However, this fuels another reason to reject $P2$.

The argument is supposed to work where Bob's decision is not relevantly different between the actual sequence (w) and a possible world identical to the actual sequence *without Black's presence*. However, in order to block Bob's alternative options (and make $P4$ true), Black actively manipulates Bob's brain *in the actual sequence* (in w). That is a relevant difference.⁹⁴ The addition of P neutralizing or manipulating parts of Bob's brain in the actual sequence makes $P2$ very implausible. The only way to get to $P2$'s truthfulness would be to assume determinism (or this

⁹⁴ One can certainly consider Pereboom's four-stage manipulation argument at that point (2022, pp. 33-39). Robb (2020, p. 4.3), himself, admits that this does seem to be a relevant difference as that it very well may be assuming determinism again:

It turns out that this example too involves blockage, as the process set up by Black has also "neutralized" the neural realization of any alternative decision... Among the reasons to worry about blockage cases is that they may, contrary to what's intended, render Jones' choice causally determined, again putting his responsibility into question. After all, the objection goes, there seems little if any difference between saying (i) that structures and processes set up in Jones' brain block off all alternatives other than choosing to vote for the Republican, and (ii) that Jones' choice is causally determined by those same structures and processes.

kind of manipulation) was compatible with free will. Therefore, this argument would be, again, question-begging against the consequence argument like the Standard-version of the Frankfurt-styled objection.

However, as I mentioned before, there have been numerous attempts at successful Frankfurt-styled cases over the years. In the next section, I am going to discuss a novel case that is, I think, more difficult to handle.

5.2.7 John Martin Fischer's New Frankfurt Argument

In §5.2.4, I explained why I do not think that the standard Frankfurt-styled objections are successful – namely, because of the Dilemma Defense. In §5.2.6, I explained why I do not think Mele and Robb's specific Frankfurt-styled objection succeeds. In this section, I will consider a case that is able to overcome those defense in an interesting and novel way.⁹⁵

John Martin Fischer wishes to offer a way of defending the claim that we can construct an indeterministic FSC [Frankfurt-styled case] in which all alternative possibilities, robust or not, are eliminated" (Fischer, 2022, p. 1189). The problem, which he admits (Fischer, 2022, p. 1190), is that

it does seem that the computer cannot 'know' with certainty at T-1 that Jones will begin to choose to vote for the democrat at T, if we assume that indeterminism is true...A mere

⁹⁵ There are some precursors to Fischer's (2022): See Michael Almeida (2016, p. 197); c.f., Lewis (2000).

human being, even assisted by a fancy computer, cannot know with certainty at T-1 that Jones will begin to choose to vote for the democrat at T. Alternative possibilities are like ants – they keep coming back.

So, what is Fischer's strategy? Simple: replace Black with another individual, one that can "know with certainty at T-1 that Jones will begin to choose to vote for the democrat at T, even in a causally indeterministic version of the case" (Fischer, 2022, p. 1190). As it was shown earlier, it is certainly difficult to imagine an indeterministic Frankfurt-styled case where the agent is responsible and free, yet has no genuine alternatives, but Fischer imagine a person with the requisite "impressive cognitive powers" (ibid): God!

God is the replacement for Black. God is a good candidate for the antagonist in the Frankfurt-story because, according to certain theistic conceptions, he has the requisite ability to know what will and would happen in indeterministic worlds (clarification: he is not contending that God is evil for this case to work!). Fischer (2022, p. 1191) continues

Although the critics of this possibility express themselves confidently, many important theistic philosophers have indeed attributed to God foreknowledge of free human behavior in an indeterministic world: Molinists and Ockhamists are examples. They argue that causal determinism would rule out human freedom, but that we are indeed free (in the sense of freedom to choose/do otherwise). Further, they contend that God has foreknowledge of this free behavior. The Molinist's central analytical device is 'middle knowledge', and the Ockhamist's is the distinction between hard and soft facts. They each put forward complex views, but they have this in common: God can have foreknowledge of free human behavior in an indeterministic context, where this knowledge is not gained from a direct gaze into the future (or some sort of crystal ball).

At this point, it will probably help to give a brief explanation on Molinism and Ockhamism as they pertain to Fischer's point here. When it comes to futuristic-will propositions (e.g., it will rain tomorrow), following Todd (2021), *will* should be treated as a universal quantifier over available futures (Todd, 2021, p. 21). Available futures should be understood as "causally possible (maximal) ways things might go from here, consistently with the past and laws (such branches will be segments of traditional abstract possible worlds)" (Todd, 2021, p. 55). If determinism is true, then there is only ever one causally possible future, and thus one available future.⁹⁶ However, if indeterminism is true, there are multiple causally possible futures. However, some theists want to maintain that God can still know *which* available future will actually be the case. So, in the context of indeterminism, on some models (e.g., Ockhamism), "even if there are multiple futures consistent with the past and the laws, there are primitive future-directed facts, and it is determinate what they are; thus only one given future-the unique actual future-is *available*" (Todd, 2021, p. 96). According to Ockhamism, even though the future is indeterminate (and, probably, no human can really know what is going to happen), there is a fact of the matter about what will happen. Todd refers to these "facts of the matter" as *primitive future-directed facts*. If God knows these primitive future-directed facts, assuming they exist, then God can know the truth value of will-propositions *even if indeterminism is true*.

Additionally, when it comes to subjunctive counterfactuals (e.g., if the match were dry, it would have lit), there is an interesting discussion similar to the one about the will-propositions. We tend

⁹⁶ Some theorists want to suggest that there are multiple *macrophysical* available futures even if determinism is true. See List, 2019b and Loewer, 2020. For the sake of simplicity here, I am putting that kind of distinction aside.

to evaluate counterfactuals based on possible worlds, and the closeness of those possible worlds to our world. If indeterminism is true, then in truly indeterministic scenarios there are *ties* in which possible world or set of possible worlds is closest to the actual world. For example, if I tossed a (truly indeterministic) coin (c), *would* it land on heads (h) or tails (t)? Well, in the indeterministic case there is a possible world where $(c \wedge h)$ is true and one where $(c \wedge t)$, and both of these worlds are equally similar to the actual world. So, what is the answer to the question? Well, on the theistic supposition like the one Fischer posits (e.g., Molinism), God would need to have a definite answer to this question. In this sort of scenario, there is a primitive modal fact of the matter that provides the definite answer. Todd (2021) and Stefansson (2018) refers to these primitive modal facts as “counterfactuals.” As Todd puts it, “That is, even if some p worlds where q and p worlds where $\neg q$ are just as similar to the actual world, there is also a further consideration, in virtue of which one of these worlds is primitively “selected”, so therefore uniquely counterfactually available” (Todd, 2021, p. 99). With the inclusion of counterfactuals to the range of modal facts, we have an explanation of how God can know contingent counterfactuals (e.g., if I were in such-and-such-circumstance, I would have stolen the money); he knows the truth value of these counterfactuals because he knows the modally primitive counterfactuals.

We are then supposed to imagine that God exists and that He can know will- and would-propositions perfectly (i.e., without error) by knowing the counterfactuals and primitive future-

directed facts.⁹⁷ So, with this background, if we plug God into the story, we can get roughly the following. At T-1, God sees Jones' neurological pattern and knows what Jones *will* do, that he will vote for the democratic candidate. However, if God believed (and thus knew) that Jones was going to vote for the Republican candidate based on Jones' prior neurological patterns at T-1, "God would have blocked him from beginning to choose to for the republican at T. So there is no way Jones could have begun to choose to vote for the republican at T" (Fischer, 2022, p. 1199). Here it is not just that God foreknows Jones' decision; it is that he has the requisite kind of foreknowledge and counterfactual knowledge to counterfactually intervene in the alternative sequence well before Jones had anything resembling a decision. It is the combination of knowing what will happen and planning for what he would do if something else would happen that allows

⁹⁷ Fischer's particular case is slightly different. He accounts for God's knowledge of counterfactual and future contingents through a bootstrapping view of God's omniscience. He wants to argue that humans can have knowledge in certain "knowledge conferring situations" (KCS's), and these situations are not (for humans) certainty-granting, but for God they are. He says,

If God believes that p and is in a KCS with respect to p, He thereby has knowledge that p. But unlike human beings, God is essentially omniscient, and so He has certainty that p, when He knows that p. How is this consistent with fallibilism about knowledge? Whereas a human being can know that p by virtue of being in a KCS, even though the evidence does not give him certainty, God must have certainty. The answer is that unlike an ordinary human being, God knows that if He believes that p, then it follows of necessity that p is true. He knows this via His self-knowledge; He knows that He is essentially omniscient. Thus, not only does God know that p, He knows it with certainty. God can thus 'bootstrap' His way to certainty in this distinctive way. We might call the approach to God's foreknowledge I am sketching, 'the Bootstrapping View'" (2022, 1194).

On Fischer's view, God knows *less* than the conception of God according to the Molinist and Ockhamist. However, I struggle to grant the possibility of this kind of bootstrapping foreknowledge and counterfactual knowledge. Consider a situation where I believe (because I am in a KCS) it will rain tomorrow, but God (being in the same KCS) does not. Here, God will know that we are both in a KCS. Given that he does not have knowledge and I do (because his essential omniscience is preventing him from doing so), he will naturally infer, and, hence, know that there is something about the fact of the matter about whether it will rain tomorrow. I do not see how this semi-fallibist concept of God's foreknowledge can avoid the kind of installment of counterfactuals and primitive future-directed facts that Ockhamism and Molinism entail. So, I will operate as if these primitive modal facts are present in any of these theories, although I acknowledge that Fischer is trying to give an account that avoids them (for Fischer's brief response to my line of argument see 2022, p. 1197-1198).

God to take away alternative possibilities. As Fischer says, “Replacing Black with God shows how the indeterministic horn of the Dilemma Defense of PAP fails” (Fischer, 2022, p. 1206).

Here, I can agree with Fischer that such a case would genuinely take out all of Jones’ relevant alternative possibilities. Thinking back to my response to the Drowning Case (5.1.3), I can agree that God, in this conception, could really take away *any* robust alternative possibilities. He could take away all options but one in this fashion (no matter how fine-grained the options were).

It will be helpful to formally list out the argument with God plugged in as the antagonist character. As I have laid the Frankfurt-styled argument, using God in this manner provides a strong reason to accept P4 is true:

P1: Jones’ choice to vote democrat would have been free in a possible world w^* identical to the one mentioned except where God was not present. (Assumption)

P2: There is no relevant difference between Jones’ decision to vote democrat in w and w^* . (Assumption)

P3: Therefore, Jones’ choice to vote democrat is free in w . (From 1, 2)

P4: There is nothing Jones can do in w that would falsify Jones’ decision to vote democrat. (Assumption)

P5: There is nothing Jones can do that would falsify Jones’ decision to vote democrat and Jones’ decision to vote democrat is free in w .

In this argument, like in the standard Frankfurt-styled argument, P5 would be a problem for premise 4 of the B-MODAL CONSEQUENCE ARGUMENT and P-MODAL CONSEQUENCE ARGUMENT. The advantage to this version of the Frankfurt-styled argument is that the Dilemma Defense does not work against it. Indeterminism is genuinely and robustly true and yet P4 is still true.

5.2.8 My Response

Going back to Fischer's argument (2022), as previously mentioned, the idea that there is nothing Jones can do in w that would falsify Jones' decision to vote democrat (P4) is powerfully motivated by the background knowledge of God's foreknowledge and counterfactual knowledge. Jones really cannot do anything other than what he actually does in this scenario because God would pre-empt Jones doing anything else. God (in this conception) clearly has the requisite ability, unlike Black's device, to ensure this. In the version of his argument that I laid out: P3 is the premise stating that Jones is free in w (the world where God is present), P3 is entailed by P1 (i.e., Jones' choice to vote democrat would have been free in a possible world w^* identical to the one mentioned except where God was not present) and P2 (There is no relevant difference between Jones' decision to vote democrat in w and w^*). We intuitively accept P1 based on a simple pre-theoretical inference (i.e., that ordinary persons are free when they make ordinary decisions). I want to grant at the outset that the conceptual inclusion of God (and his perfect foreknowledge and middle knowledge) complicates the discussion of free will and can be difficult to reconcile. In this section, I want to offer a couple of responses against P2.

My first response takes inspiration from Steward (who is replying to Hunt, 2000 and Fischer, 1999), when she notes that:

The entanglement of modality with the characterization of actuality is, after all, not an unfamiliar phenomenon in philosophy. Where substances are concerned we are used to the entanglement. Suppose one is told to imagine a world that is just like the actual world, except that in the imagined world Tony Blair could have turned into a teapot at any time. Everything actual, we are assured, is just the same, and Tony Blair never in fact turns into a teapot. We are only to imagine a world in which it is true that he could have done. It is only a modal fact that is different. But it is likely immediately to be objected to the coherence of this world-description that, the modal facts having been tampered with in this way, we no longer have something that counts as a world that contains Tony Blair at all. For surely anything capable of turning into a teapot at any moment could not actually be Tony Blair; Tony Blair, one might think, is essentially a human being, and nothing that is essentially a human being is capable of turning into a teapot at any moment. In this case, tampering with the modal facts is enough to alter the intrinsic description we are prepared to supply of the world itself. The lesson is that what can be conceived of as existing in a world is simply not independent of what can happen there... Tampering with the available alternate possibilities may affect the question whether something that is genuinely the action of an agent may be said to have occurred or not, in a given imaginary case. (Steward, 2013, 188).

Her point about Tony Blair having the power to turn into a teapot (even if he never actually does) is instructive. Imagine Tony Blair under normal circumstances (where he is not magical): if I were to claim that Tony Blair *cannot* run from London to Jerusalem in 10 minutes, and I were to back

this claim up with the reason that he is an ordinary human person (and not a magician) and people do not run fast enough to make it that distance and time, then my claim is clearly an intuitive inference.⁹⁸ However, once I introduce the modal fact assumption that Tony Blair actually has the power to turn into a teapot whenever he wants (he just never has done it) to my argument (about Tony Blair's inability to run to Jerusalem in 10 minutes), then my argument is not clearly intuitive at that point. We would want to account for this new property of Blair's before we assess his traveling abilities. Perhaps, for example, Tony Blair *is a magician* or a miracle worker of sorts. A magician might be able to run from London to Jerusalem in 10 minutes, so, perhaps, Blair could too.

We would not normally pause in our judgement of Blair's ability to run to Jerusalem in 10 minutes, but we normally do not even consider the possibility that he can turn into a tea-pot either. We would never consider the plausibility of the claim that he is a magician (which would intuitively make the further claim- that he can travel from London to Jerusalem in 10 minutes- plausible). However, we have to explain his ability to turn into a teapot *somehow*. So, the modal fact that he can turn into a teapot can upend all sorts of intuitive judgements about Blair. When we tamper with the modal facts, it can have an effect on our perception of an actual person's powers, and, theoretically, their free will.

Adding God (in the conception described by Fischer) to the situation adds quite a lot of metaphysical and modal import to the situation. In the possibility where God is absent (w^*), Jones

⁹⁸ This example is similar to one Van Inwagen uses in (2004, p. 349).

could do all sorts of things differently: he could have pushed the button for the democratic candidate faster or pushed the button while yelling “*Viva la democrat!*” at the top of his lungs. He presumably cannot do any of those things in the situation with God (w). But, imagining w , we have to imagine not only Jones and God, but these primitive modal counterfactuals and future-directed facts as part of w as well. This, I suggest is a relevant difference because it seems possible to, additionally, run a version of the modal consequence argument against this variety of theism and modality (or foreknowledge).

For example Patrick Todd and John Martin Fischer (Todd & Fischer, 2015, p. 2) write an example fatalistic argument for Incompatibilism using future-directed facts in the following manner:

- (1) You had no choice about: it was true 1,000 years ago that you would purchase the book at t .
- (2) Necessarily, if it was true 1,000 years ago that you would purchase the book at t , then you purchase the book at t . So,
- (3) You had not choice about: purchasing the book at t .

Although the fatalist argument may be different than the consequence argument, we can easily re-write this argument in the form of the Modal Consequence Argument:

- (1) There is nothing anyone (including you) can do that would falsify a proposition, P , expressing that it was true 1,000 years ago that you would purchase the book at t . (Fixity Principle, or NP)
- (2) Necessarily, if it was true 1,000 years ago that you would purchase the book at t (P), then you purchase the book at t (B). ($\Box (P \rightarrow B)$).

- (3) Therefore, there is nothing anyone (including you) can do that would falsify a proposition expressing that you purchase the book at t . (Or, NB, from 1 and 2, and Beta-Best).
- (4) If (3) is true, then you can have free will with relation to your purchasing the book at t
(ALTERNATIVE-POSSIBILITIES)
- (5) Therefore, you can have free will with relation to your purchasing the book at t

In this argument, (2) is fairly uncontroversial (it is really an expression of what Van Inwagen calls Rule Alpha, see (1989)). The most prominent objection would be that (1) is false because P depends on you in some sense (Todd & Fischer, 2015, pp. 4-5). If that is the case, then this argument does not work. However, this is very controversial, and this makes using these sorts of modal facts *in the Frankfurt-case* unhelpful. If, in discussing the Frankfurt-case, we end up just discussing a different version of the Modal Consequence Argument, it lessens the importance of such an objection. Just like adding magical properties to Tony Blair makes us suspend or reject all sorts of ordinary judgements about Blair, we should likewise at least suspend our intuitive judgment that Jones is free in this case. (And, if the new modal consequence argument is sound, then we can simply reject the claim altogether).

I do think there is an additional reason to reject P2 though. The second response to P2 is that the difference-making relationships are asymmetrical in Fischer's scenario, which plausibly makes a relevant difference. In w (i.e., the possible world where God preempts Jones' decision) we can clearly see the following counterfactual is true:

(C1): If Jones had decided to vote democrat, then he would have voted democrat.

However, it is false (because of God's preemption) that,

(C2): If Jones had not decided to vote democrat, he would have not voted democrat.

Here, only the positive counterfactual (C1) is true; the negative counterfactual (C2) is false.

Furthermore, consider the fact that the following is true,

(C3): If God had decided for Jones to vote democrat, then Jones would have voted democrat.

However, (unlike C2) the following is true,

(C4): If God had not decided for Jones to vote democrat, then Jones would not have voted democrat.

So, both counterfactuals (the positive and negative counterfactuals) hold true for God's relationship to Jones' voting or not voting democrat. This could be a relevant difference between w and w^* . If we are looking for an explanation for *why* Jones' voted democrat, a natural explanation would be that God *decreed* he would do so. God's decision makes more of a difference than Jones decision in w . In w^* (the possible world without God), there is no difference-making cause or explanation of Jones' voting other than Jones' deciding to vote democrat. Without God's preemption we would assume the counterfactual 'if Jones did not decide to vote democrat, he would not have voted democrat' is true. And, if one assumes a difference-making view of causation, in w , God would be the cause of Jones' action, not Jones. While in w^* , on a difference-making view of causation, Jones would be the cause of his own action. The metaphysical import of God (in Fischer's hypothetical conception) would alter the causal facts of what caused Jones to vote democrat in w *in the actual sequence*. Thus, we should infer that P2 is false.

Conclusion

This chapter has discussed ALTERNATIVE-POSSIBILITIES and OPTION-BASED ALTERNATIVE-POSSIBILITIES, and two objections to Premise (4) of the B-MODAL CONSEQUENCE ARGUMENT: The first was from John Maier, who argued that it is possible for agents to have multiple options in deterministic worlds, but not options to falsify the conjunction of the past and the laws. The second objection was the Frankfurt-styled argument, which argued that because a Frankfurt case is successful, then Premise (4) of the B-MODAL CONSEQUENCE ARGUMENT is false. I argued that both of these types of objections fail.

These two objections are important to address for the proponent of the B-MODAL CONSEQUENCE ARGUMENT (and the proponent of the P-MODAL CONSEQUENCE ARGUMENT) because Premise (4) - that assume a form of ALTERNATIVE-POSSIBILITIES- is essential for the argument's success. The first three chapters of this thesis were setting the stage for the main argument that I would defend (i.e., the B-MODAL CONSEQUENCE ARGUMENT). Chapter Four defended the argument's second premise, which was the key Fixity Principle. This chapter defended the other controversial premise in the argument. Having defended these two premises (and the Beta Rule in Chapter Two), I have shown that this is a powerful and sound argument. The conclusion of this argument is that by necessity, if non-reductive physicalism is true, then no one can have free will (6). So, if (6) is a sound and convincing conclusion, what are we to do with this? Well, in the next chapter, I will show there are various implications of (6) that we could infer.

CHAPTER SIX: INTERACTIONIST DUALISM AND FREE WILL, FREE WILL IMPOSSIBILISM, OR FREE WILL REVISIONISM

Introduction

In this concluding chapter, I will review what I have accomplished in the thesis so far and consider some possible implications of, or reactions to, the B-MODAL CONSEQUENCE ARGUMENT. As I see the landscape of possibilities, this argument has three clearly available options regarding free will we can consider next:

1. Interactionist dualism and free will are both true.
2. Free will is impossible in the actual world (because non-reductive physicalism is true in our world).
3. We ought to revise the conditions of free will (i.e., revisionism is true).

Which of these remaining positions is the correct view is not something I am going to argue conclusively in this chapter. My goal is to show that all of these views are live options, given the plausibility of the B-MODAL CONSEQUENCE ARGUMENT.

I am going to begin this chapter by reviewing Chapters One through Five (§ 6.1). Then, in § 6.2, I am going to consider the first possible reaction to the B-MODAL CONSEQUENCE ARGUMENT, which is to explore the possibility that dualism and free will are both true (or that free will requires something like dualism). Afterwards, I will explore the possible reaction that non-reductive physicalism *is true* and thus free will is impossible (§ 6.3). Then, I will discuss the possibility that the correct reaction is to revise the conditions for free will (§ 6.4).

6.1 Review

The thesis started by asking: Imagining God has two shelves of ingredients-physical and non-physical, would God have to reach the second shelf (i.e., with non-physical ingredients) to create free creatures? In §1.3, I pointed out that an appropriate conception of *physical ingredients* would be suggesting physicalistically acceptable ingredients. On this conception, physical ingredients would be either the fundamental properties expressed by a completed physics (where this completed physics does not posit phenomenal consciousness as a distinct fundamental property) or genuine non-fundamental properties that are (i) dependent on these fundamental properties and (ii) consistent with the closure doctrine. For example, a property like “having spin” or “weighing .2 grams” would be good candidates for physical ingredients. Likewise, an appropriate conception of *non-physical ingredients* would be non-physicalistically acceptable objects and properties. These include properties like “being a ghost.”

Additionally, I defined free will in the following way (in §1.1):

FREE WILL: A person P has free will with regard to ϕ -ing *if and only if* (I) P has the ability to consciously choose an outcome ϕ (i.e., form an intention to ϕ or make a decision to ϕ) where ϕ is distinctly caused by P’s conscious intention, and (II) when P made a decision to ϕ , P could have chosen $\neg \phi$.

I argued the most natural way for the physicalism to account for free will would be to accept MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY, which (in conjunction with GLOBAL SUPERVENIENCE and NOMOLOGICAL CLOSURE) is a position I refer to as non-reductive physicalism. Non-reductive physicalism is *prima facie* compatible with free will. However, this compatibility needs to be tested, which is why I considered an argument that appeared relevant to this question.

In Chapter Two, I explained the consequence argument, which is an argument for the conclusion that free will and determinism are incompatible. This argument seems relevant for non-reductive physicalism as both determinism and non-reductive physicalism posit something (which is seemingly not the person herself) as necessitating one's decisions and actions. In the case of determinism, the necessitating state is the remote past (with the laws of nature), while in the case of physicalism, it is microphysics. Here, I showed (what I took to be) the best version of the consequence argument: The P-MODAL CONSEQUENCE ARGUMENT (named after Alexander Pruss). The most common objection to the P-MODAL CONSEQUENCE ARGUMENT, and all modal consequence arguments, is the McKay and Johnson (1998) objection to Rule Beta. In this chapter, I showed how Pruss' reply to McKay and Johnson is successful and arguably strong enough to fend off any further objections along the same lines. I also defended this argument against another structural criticism from McKenna and Pereboom (2013). My main goal in this chapter was to show the P-MODAL CONSEQUENCE ARGUMENT is a valid argument and BETA-BEST was well justified.

This allowed me in Chapter Three to re-orient the structure of the P-MODAL CONSEQUENCE ARGUMENT so that an argument like it can be applied to non-reductive physicalism. The resulting B-MODAL CONSEQUENCE ARGUMENT claims that, because every decisions a person makes is (I) realized (or grounded) by a microphysical state S and (II) S is nomologically closed, therefore if non-reductive physicalism were true, no person could have a free will. According to non-reductive physicalism (I) S (the microphysical state realizing one's decision) metaphysically entails a person P's decision and (II) P cannot do anything that would make S not obtain (because of NOMOLOGICAL CLOSURE). From these two truths, one can use BETA-BEST to argue that, therefore, no one can do anything other than what they actually do. As long as ALTERNATIVE-POSSIBILITIES is true, then no one can have a free will. Thus, if non-reductive physicalism is true, no one can have free will. This

is an important claim, and I acknowledged that many critics would likely object to one of two key premises of the argument, either the claim that no one can do anything to falsify S (The fixity of the microphysical configuration principle) or ALTERNATIVE-POSSIBILITIES.

So, in the next two chapters, I discussed these two controversial aspects of the B-MODAL CONSEQUENCE ARGUMENT. In Chapter Four, I defended the FIXITY OF THE MICROPHYSICAL CONFIGURATION, and argued that the combination of NOMOLOGICAL CLOSURE and MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY plausibly justify this fixity principle. In this argument, I explained two ways in which the physicalists could understand mental causation: the Subset view (Wilson, 2011) and the intra-level causation view (List, 2019). I argued that the intra-level causation view is perfectly compatible with the FIXITY OF THE MICROPHYSICAL CONFIGURATION. However, I acknowledged that the subset view would potentially be a challenge to the FIXITY OF THE MICROPHYSICAL CONFIGURATION. I subsequently argued that the subset view has a significant problem in its reliance to a counterfactual account of causation, which I argued would be under-sensitive with regards to mental causation (i.e., the problem of additivity), and, thus is not plausibly a view of mental causation.

In Chapter Five I considered the plausibility of ALTERNATIVE POSSIBILITIES principle in the argument. I began by introducing a modal account of abilities from John Maier (2015), and argued that a better way to conceptualize the ALTERNATIVE POSSIBILITIES principle was OPTIONS ALTERNATIVE-POSSIBILITIES, and defended this principle against the Robustness Objection, and Maier's argument against the consequence argument. Further, I considered the objection to the B-MODAL CONSEQUENCE ARGUMENT from Frankfurt-styled Objections, and argued that none of them were successful.

So, at this point in this thesis, I have defended the three controversial aspects of the B-MODAL CONSEQUENCE ARGUMENT (i.e., (1) BETA-BEST, (2) FIXITY OF THE MICROPHYSICAL CONFIGURATION, and (3) ALTERNATIVE-POSSIBILITIES). Now, it is time to consider some of the implications and applications of the B-MODAL CONSEQUENCE ARGUMENT.

The B-MODAL CONSEQUENCE ARGUMENT provides an argument for the claim that non-reductive physicalism and free will are necessarily incompatible. In other words, if God is trying to create a free creature (e.g., a person with free will) in a possible world where non-reductive physicalism is true, he will be missing some ingredient. If this is true, what options do we then have concerning the question of whether we have a free will in the actual world?

6.2 Interactionist Dualism

The B-MODAL CONSEQUENCE ARGUMENT would seem to imply that some sort of non-physicalism is necessary for free will. Put another way, it would seem to imply that if free will exists in the actual world, there is some non-physical ingredient in the actual world. Of the most predominant non-physicalist theories, interactionist dualism is one the most prominently discussed option in the literature; therefore, here, I am going to consider the possibility that interactionist dualism is (for some reason) necessary for free will.⁹⁹

⁹⁹ There are other non-physicalist possibilities that one could explore. For example, perhaps some contemporary form of idealism (Rasmussen, 2023) or panpsychism (Goff, 2017) is compatible with and sufficient for free will. My comments on considering dualism are not meant to convey that other non-physicalist are impossible. Additionally, there are dualistic theories that are not interactionist theories. A dualistic theory denying that phenomenal states are causally affecting the world would be a variety of epiphenomenalism, like in William Robinson (2019). Here, I am going to focus on theories where the phenomenal interacts with the physical (i.e., interactionist dualism) because I take it that epiphenomenalism is incompatible with free will.

Following David Chalmers (Chalmers, 2010, p. 126), I will think of interactionist dualism in the following manner:

[Interactionist Dualism:] holds that microphysics is not causally closed and that phenomenal properties play a causal role in affecting the physical world. On this view...physical states cause phenomenal states, and phenomenal states cause physical states. The corresponding psychophysical laws run in both directions. On this view, the evolution of microphysical states is not determined by physical principles alone. Psychophysical principles specifying the effect of phenomenal states on physical states also play an irreducible role.¹⁰⁰

In Chalmers' description, interactionist dualism is mainly about two separate principles. Firstly, interactionist dualism necessarily denies NOMOLOGICAL CLOSURE-the view that microphysics is closed- and, secondly, accepts the view that phenomenal states are not ontologically or causally identical to microphysical states (i.e., MENTAL ONTOLOGICAL AND CAUSAL IRREDUCIBILITY).¹⁰¹ Following Chalmers, I categorize any view that meets these two criteria as an interactive dualist view.

Aside from these two aspects, dualists will disagree about several issues; For example, dualists disagree whether there are non-physical objects and substances (e.g., souls, ghosts, God) or simply

¹⁰⁰ Chalmers refers to this as Type-D Dualism (D is for Descartes).

¹⁰¹ There could be other mental states that are not reducible either on this view.

properties, how phenomenal properties come into the universe (e.g., emergentism theories), and, especially, how the non-physical stuff (e.g., phenomenal states) causes or affects the physical world.

The reason I am considering dualism, though, is that it plausibly offers some missing ingredient for free will that physicalism does not. So, what ingredient would dualism provide to the metaphysical picture of free will that physicalism is lacking? I think the simple answer is that according to dualism, phenomenal states can causally impact microphysical states in a manner contradictory to NOMOLOGICAL CLOSURE.

However dualists can explain this causal impact in multiple ways. One possible causal story is to suggest that dualism provides a kind of non-physical *sui generis* energy, force, or power that is not available in the physicalist picture (Hart, 1988; Stapp, 1993; Pitts, 2020; Plantinga, 2007; 2011; Dimitrijevic, 2020; Wilson, 2021). On this view, there is generally some incorporation of an interpretation of quantum physics or conservation laws that is compatible with phenomenal states (or other non-physical entities) influencing the physical goings-on. Another, related but slightly distinct, possible answer would be to suggest that dualism offers unique causal explanations that physicalism lacks (Lowe, 2000; Gibb, 2010; 2013; 2015a; 2015b; 2015c). For example, on E.J. Lowe's picture, there are two fundamental causal *relata*, events and states (or facts). So, phenomenal states can cause physical states in a way distinct from physical events.

Of course, the mere mention of dualism will invoke the objection that we are adding to free will some "spooky" or extravagant properties, and, thus, dualistic free will "is not credible given our best physical theories" (Pereboom, 2011, p. 407). So, in order to understand the plausibility of dualistic metaphysical ingredients, it will be helpful to first review a prominent argument against dualism (and dualistic free will) which is the Argument from Conservation Laws.

6.2.1 The Argument from Conservation Laws

Conservation laws are laws concerning quantities in physical systems; the laws of thermodynamics for example, which governs energy. A refrigerator, for example, is thought to be a conservative system because, although it dispels heat from its inside container, it correspondingly uses outside electricity to cool the inside container, meaning that the broader system (i.e., the refrigerator and the electricity system) does not create, nor destroy energy.¹⁰² Many physicalist philosophers have argued that conservation laws disprove interactive dualism by proving NOMOLOGICAL CLOSURE. For example, in David Papineau's famous paper "The Rise of Physicalism," he recounts that for a long time he thought it was obvious the closure principle "follows from the fact that physics can be formulated in terms of conservation laws" (Papineau, 2001, p. 185).

To that end, Daniel Dennett argues that once one simply thinks about the dualistic hypothesis, one realizes that the mind causes physical states without a causal *relata* or process that is, in any shape or form, physical (e.g., there is no energy transference), and this by itself should be problematic (1991, p. 35). As Dennett notes, a basic premise of physics is that physical change requires an expenditure of energy (i.e., conservation laws), which is why, for example, perpetual motion machines are impossible (ibid). One can easily, then, argue that conservation laws should eliminate the possibility of dualism in the same manner they eliminate perpetual motion machines. It is conservation laws that conflict with dualism and, leads philosophers like Dennett to say, "[this]

¹⁰² Physicalists have typically justified NOMOLOGICAL CLOSURE by appealing to conservation laws. For example, in David Papineau's famous paper on nomological and causal closure, he claims he always assumed the principle "follows from the fact that physics can be formulated in terms of conservation laws" (Papineau, 2001, p. 185)

confrontation...has been endlessly discussed since Descartes's own day, and is widely regarded as the inescapable and fatal flaw of dualism" (ibid).

Or, as Mario Bunge (1980, p. 17) writes,

Dualism violates conservation of energy. If immaterial mind could move matter, then it would create energy; and if matter were to act on immaterial mind, then energy would disappear. In either case energy—a property of all—and only concrete things would fail to be conserved. And so physics, chemistry, biology, and economics would collapse. Faced with a choice between these “hard” sciences and primitive superstition, we opt for the former.

The reasoning here is fairly clear. Dennett and Bunge are arguing that if one accepts basic conservation laws, then one should reject interactionist dualism. It seems that many philosophers in the twentieth century (especially) have endorsed this argument.¹⁰³

In order to examine the argument formally, we can utilize Sophie Gibb's definition of the principle of conservation (Gibb, 2015, p. 117),

¹⁰³ John Searle, additionally, says “[a]ll forms of substance dualism inherit Descartes' problem of how to give a coherent account of the causal relations between the soul and the body, but recent versions have an additional problem. It seems impossible to make substance dualism consistent with modern physics. Physics says that the amount of matter/energy in the universe is constant, but substance dualism seems to imply that there is another kind of energy, mental energy or spiritual energy, that is not fixed by physics. So if substance dualism is true then it seems that one of the most fundamental laws of physics, the law of conservation, must be false" (Searle, 2004, p. 42). Jerry Fodor, likewise, remarks, “[t]he chief drawback of dualism is its failure to account adequately for mental causation. ... How can the nonphysical give rise to the physical without violating the laws of the conservation of mass, of energy and of momentum?" (Fodor, 1994, p. 25).

[CONSERVATION]: Every physical system is conservative or is part of a larger system that is conservative (where a system is conservative if its total amount of energy and linear momentum can be redistributed, but not altered in amount, by changes that happen within it).

In this principle, the idea is that for any physical system P, the only way for amount of energy in P to change is if P is part of a large physical system P*. For example, the runner's body might have a certain amount of energy (in a system P), but that can change if she drinks Gatorade (which would now be included in the larger system P*). Additionally, the distribution (i.e., the spatial-temporal properties of the energy including their momentum) of the energy in P can only change from within P, unless P is part of a large physical system P*. For example, the molecules in a cup of hot coffee reposition themselves constantly as the entropy increases and lowers the temperature of the beverage to the room's temperature (which could count as the larger system, as the room temperature will impact the temperature of the beverage). Also, note that CONSERVATION suggests that *every* physical system is conservative, and the universe is thought to be a physical system, so, therefore, the universe must be a conservative system as a whole.

With CONSERVATION, we can form the following argument formal argument:

- (1) CONSERVATION is true
- (2) If CONSERVATION is true, then dualism is false.
- (3) Therefore, dualism is false.

I take this argument to be the heart of what Dennett and Bunge are arguing.¹⁰⁴ It is simple, but the conservation laws are supposed to disprove dualism because phenomenal states (or souls or non-physical minds) do not have any (distinct) energy (or other relevant physical quantity or property) associated with them.

However, dualists are not very persuaded by this argument. Alvin Plantinga, a notable dualist, for example, goes so far to suggest that this “objection, therefore, is wholly without force. It’s not that it gives one some reason, perhaps only a weak reason, for rejecting dualism; it provides no reason at all” (2007, 127). They are not very persuaded, generally, because there are a few distinct strategies of objecting to the argument. All of these strategies will help to explain the different ways in which dualism can contribute to the explanation of what constitutes free will (the necessary conditions or ingredients).

6.2.2 Three Dualistic Responses

There are different ways in which dualists can argue against CONSERVATION (1) or the claim that CONSERVATION entails that dualism is false (2).

Firstly, one might argue against the idea that CONSERVATION would entail dualism is false (premise 2) by pointing out that, as Chalmers notes, our current physics does not specify exactly *how many* fundamental forces there are (we currently have four); thus, “an extra mental field would be a

¹⁰⁴ David Papineau (2000; 2001) argues inductively in a slightly different way. I will bring up his argument shortly.

reasonable extension of existing physical theory” (2010, 128). CONSERVATION is not applicable to subatomic particles or atomic nuclei, where we deal with these fundamental forces.

Papineau, one of the most famous proponents of the Argument from Conservation Laws, agrees that this sort of response is generally open for the dualist to take, but there is an obvious price. Papineau argues that most of the forces we observe do seem special or novel at first. However, the vast majority of these ultimately reduce down to the fundamental forces we currently postulate. He gives the example that friction might seem *prima facie* special in some sense, but upon analysis reduces to more conservative forces (2010, 28). From this he argues inductively that any perceived mental or phenomenal *force* is likely to reduce down to a non-mental force like friction. However, he says, “[this] is of course not a knock-down argument. Vital or mental forces could figure among the fundamental forces of nature, even if they are only generated in the special circumstances associated with life or sentience” (ibid). He argues that this is very unlikely though.¹⁰⁵

Second, and similar to the first possible response, a more prominent argument against the second premise (that If CONSERVATION is true, then dualism is false) is that CONSERVATION is not a law that is applicable in quantum mechanics, and phenomenal states have a role in determining microphysical states (Collins, 2008). According to the standard, Copenhagen, interpretation of quantum mechanics, when we apply the Schrödinger equation to microphysical systems (e.g., a group of electrons), we describe the system using a wave function according to which the physical

¹⁰⁵ On this strategy, dualists would point out that our physics will change in the future, and it is conceivable that it will include a kind of dualism (via a mental force). However, if the argument is inductive, it will arguably hinge of the prior probabilities of both dualism and physicalism.

entities are in a superposed state (e.g., in a superposition of two different positions in an experiment).¹⁰⁶ However, we do not directly observe the superposition of a particle (or group of particles) because the superposition is not locatable in classical space-time. When we measure the wave function, it *collapses* into a classical time-space quantity that we can observe, and this collapse mechanism can be interpreted friendly to dualism. As Chalmers (2010, pg., 128) notes:

The collapse dynamics leaves a door wide open for an interactionist interpretation. Any physical nondeterminism might be held to leave room for nonphysical effects, but the principles of collapse do much more than that. Collapse is supposed to occur on measurement. There is no widely agreed definition of what a measurement is, but there is one sort of event that everyone agrees is a measurement: observation by a conscious observer. Further, it seems that no purely physical criterion for a measurement can work since purely physical systems are governed by the linear Schrödinger dynamics. As such, it is natural to suggest that a measurement is precisely a conscious observation and that this conscious observation causes a collapse.

This is precisely the way the Nobel Prize winning physicist Eugene Wigner (1961, pg., 168) took the collapse mechanism to be:

¹⁰⁶ A Superposition is a mathematical description an electron that is not describable in classical 3 dimensional space. In an experiment where an electron is passing through an apparatus with two possible exists (*h* and *s*), Albert describes like this: “Electrons passing through this apparatus, in so far as we are able to fathom the matter, do not take route *h* and do not take route *s* and do not take both of those routes and do not take neither of those routes; and the trouble is that those four possibilities are simply all of the logical possibilities that we have any notion whatever of how to entertain!” (Albert, 1992, p. 11). There are various interpretations of what this wave-function is supposed to be metaphysically. Some philosophers take this to be a genuine sort of metaphysical object, see Albert and Ney (2007).

Until not many years ago, the “existence” of a mind or soul would have been passionately denied by most physical scientists. . . . [There are] several reasons for the return, on the part of most physical scientists, to the Spirit of Descartes’ “Cogito ergo sum”. . . . When the province of physical theory was extended to encompass microscopic phenomena, through the creation of quantum mechanics, the concept of consciousness came to the fore again: it was not possible to formulate the laws of quantum mechanics in a consistent way without reference to consciousness.

If one does accept quantum mechanics (and the collapse mechanism in particular) interpreted in this way, then the second premise (that if CONSERVATION is true, then dualism is false) is not true. Phenomenal states can have a role simply at a different level than what our conservation laws govern. As J.M. Schwartz et al, write, “Our willful choices enter neither as redundant nor epiphenomenal effects, but rather as fundamental dynamical elements that have the causal efficacy that the objective data appear to assign to them” (Schwartz, 2003).¹⁰⁷

There are two notable problems with this application of dualism. One is that it is clearly true that every theory of quantum mechanics is controversial, so it may be tenuous to stake interactionist dualism and free will on one particular quantum theory (or one family of theories).¹⁰⁸ For example,

¹⁰⁷ Furthermore, the dualists can suggest, as Chalmers opines, that “if one were to design elegant laws of physics that allow a role for the conscious mind, one could not do much better than the bipartite dynamics of standard quantum mechanics” (Chalmers, 2010, pg., 129).

¹⁰⁸ There are going to be other theories of quantum physics that would be compatible with this sort of dualistic collapse mechanism. Notably, the GRW theory would, in theory, be. Henry Stapp (1993) and Alvin Plantinga (2011) both have tried to elaborate further on the GRW theory and the dualistic interpretation.

if, by some manner, an Everitteen or Bohmian theory of quantum physics was confirmed, then this view of free will would be false. The second is that some philosophers (e.g., Goff, 2019) have suggested that even if consciousness does serve to collapse the wave-function, this does not give us the proper sort of control or causal affectedness we want for free will. Dualists will need to explain how collapsing the wave function can affect, for example, neural events and decisions. But, in principle, it is possible that small-scale changes in the brain at the quantum level could affect neural events and such. This would not currently be testable, but it should be something that is theoretically testable.¹⁰⁹

Third, another argument that dualists have used is to suggest that CONSERVATION is simply false (premise 1) because our conservation laws should be thought of as local governing principles, and not as universal laws holding for the entire universe (Collins, 2008; Pitts, 202). J. Brian Pitts has argued that our equations for conservation laws are *local* and there is not a single equation that can describes the energy flow for the whole universe (2021, pp., 567-586). Hence, he notes a localized “conservation can hold perfectly well in astrophysics and refrigerators, even if interactionist dualism is true’ (567). What the dualist would want to suggest is that “conservation holds in most times and places, but fails in a few, namely, where souls [or phenomenal states] act on brains” (Pitts, 2021, p. 569).¹¹⁰

¹⁰⁹ Barrett (2002) notes that this actually would be a theoretically possible way to test whether a person is conscious or not, or whether a robot has consciousness, which is an interesting implication of this view. Also, this has similarities to Robert Kane’s view of free will (1996), although he rejects the dualistic aspects.

¹¹⁰ Alternatively Plantinga (2007) suggests that conservation laws do not hold locally, but do hold globally.

Here the dualist can naturally appeal to the possibility of a Maxwell Demon to illustrate how this could physically work. To discuss Maxwell's Demon, it may be helpful to briefly characterize the two relevant laws of nature for this thought experiment. The first law of thermodynamics states (roughly) that the total energy of an isolated system is always constant. The second law of thermodynamics states (roughly) that usable energy in an isolated system always decreases. The way the second law is often defined is in terms of a concept known as entropy, which is a statistical concept measuring a system's disorder. When molecules are increasingly in disorder, the usable energy of the system decreases. A physical system is always moving toward a state of equilibrium, which is the state where the entropy is at its highest and there is no useable energy. This process, for example, is why a cup of hot coffee and a cup of iced coffee (in the same room) will eventually both be at room temperature even though they started at different temperature states. The second law of thermodynamics is particularly interesting because it is a statistical law. Just like the most likely dice role (based on the greatest realizable subset of dice rolls) is a 7 (when rolling a pair of dice), a physical system always is statistically most likely to have its entropy increase.

Now, we can consider the thought experiment. The physicist James Clerk Maxwell considered the second law of thermodynamics in the late nineteenth century, he wrote (in prior letters, and then in his 1871 *Theory of Heat*) an interesting thought experiment later called "Maxwell's Demon." In this experiment, a demon decreases the entropy of an isolated system (which is not supposed to happen under the 2nd law). Although in his 1871 version, he simply referred to the being a "finite being". Here is how he talked about this demon:

But if we conceive a being [a demon] whose faculties are so sharpened that he can follow every molecule in its course, such a being, whose attributes are still as essentially finite as our own, would be able to do what is at present impossible to us. For we have seen that the

molecules in a vessel full of air at uniform temperature are moving with velocities by no means uniform, though the mean velocity of any great number of them, arbitrarily selected, is almost exactly uniform. Now let us suppose that such a vessel is divided into two portions, A and B, by a division in which there is a small hole, and that a being, who can see the individual molecules, opens and closes this hole, so as to allow only the swifter molecules to pass from A to B, and only the slower ones to pass from B to A. He will thus, without expenditure of work, raise the temperature of B and lower that of A, in contradiction to the second law of thermodynamics. (Maxwell, 2011, pp. 308-309)

From this passage, readers are supposed to be able to conceive of a demon sitting on top of a box, which has two chambers and a small door connecting the chambers. The first chamber (B) has faster molecules because it is warmer, and the second chamber (A) has slower molecules because it is colder. This demon uses his advanced knowledge to only open the door at particular times in order to transition faster molecules into one side (B), and the slower ones into (A). By doing this, the demon can decrease entropy-the equilibrium of probability states of the molecules- to keep one box warmer or faster than the other. In other words, the demon can reverse the asymmetry of the second law of thermodynamics and always keep one room hotter than the other.

What is particularly interesting about the idea of this demon is that, if it had these sorts of powers, it could effectively *redistribute* the energy in an isolated system without breaking the first law of thermodynamics. The demon does not pump in new molecules into the whole box structure, nor does he take any out. *He merely redistributes them.* However, this redistribution does violate CONSERVATION because the demon and the box together are not conjointly an isolated system. It violates CONSERVATION because, as Dejan Dimitrijevic notes, “we have to accept that the demon is not a part of the observed physical system and that its intelligence cannot be described by the

known physical laws. Accordingly, the Second Law can be applied neither on the demon, nor on the combined system gas-demon” (Dimitrijevic, 2020, p. 12). The second law of thermodynamics is merely a statistical law showing the probability state of certain physical systems (e.g., the molecules in a cup of coffee). Certain physical systems would be conservative on the condition they are isolated or part of a bigger physical isolated system.

This demon can be thought of as an analogy for dualistic mental causation. There is nothing clearly contradictory or inconceivable about Maxwell’s Demon redistributing the energy, so why should there be anything contradictory or inconceivable about the mind redistributing energy in, say, neural processes. By conceiving of mental causation as a power to redistribute energy, all the mind would need to do is “bring about small-scale, local correlations of neural processes, hardly distinguishable from fluctuations of physical quantities, in order to produce large-scale physical effects” (2020, 10). It is true that CONSERVATION would fail, but CONSERVATION would arguably fail in Maxwell’s thought-experiment as well.¹¹¹

All three of these responses demonstrate that Dennett and Bunge’s Argument from Conservation Laws is not straightforward and compelling.¹¹² Dualists do have ways to responding to this

¹¹¹ Like Wigner’s theory that consciousness collapses the wave-function (which would be compatible with this suggestion), this proposal would be, in principle, empirically testable. However, it is not something that we can currently test in physics or neuroscience.

¹¹² A fourth response is to appeal to a different sort of causation altogether for mental causation (Gibb, 2015). E.J. Lowe (2000), for example, wants to argue for a difference between event and state causation, and then postulate mental causation as state (or fact) causation. This is an interesting possibility, but I think it is much more plausible when paired with one of the responses mentioned.

argument. Of course, there are other arguments against dualism.¹¹³ However, this objection is important because (i) it is a very popular objection against the plausibility of dualism, and (ii) the dualistic responses help us see what potential extra ingredient dualism adds that physicalism does not. In the next section, I will discuss (ii) further.

6.2.3 The Dualistic Ingredient for Free Will

Dualists do not necessarily need to appeal to souls or soul-ish stuff (e.g., Kant's transcendent self) to explain free will. Dualists need to be able to explain how phenomenal states can influence physical states. If any of these responses to the Argument from Conservation Laws succeeds, then the dualist will have a way in which to do that.

¹¹³ Notably, David Papineau argues inductively that “there is no direct evidence for vital or mental forces. Physiological research reveals no phenomena in living bodies that manifest such forces. All organic processes in living bodies seem to be fully accounted for by normal physical forces” (Papineau, 2001, p. 27). This argument, though, is not completely detached from the Argument from Conservation Laws. At a certain point, all physical movements will be caused by brain activity, of which neurons are going to be central. And, when we get to neurons, and their behaviors, we will be thrust back into quantum physics. It is an open question whether or not dualistic theories of the varieties discussed in this section are successful. However, the mechanism of the collapse of the wave-function, and way in which it could have macrophysical effects, would count as relevant data to be explained (on certain theories of quantum physics). The only way for Papineau's argument to work is if the prior probability being assumed here discounted these possibilities, which is related to the Argument from Conservation Laws. There is also a family of arguments inspired from the work of Benjamin Libet (1985; 2004) and Daniel Wegner (2002). These arguments seek to show that recent work with EEG (Libet, 1985; Schlegel, et al., 2015) and fMRI (Soon, et al., 2008) technology has shown us that actions (of the kind studied in these particular experiments) do not have conscious intentions among immediate causes, and thus, (generalizing) all actions do not have conscious intentions causing them. These arguments work because of various experiments where patients perform some action (e.g., hitting a button) at a certain time, and then these technologies reveal a certain neural build-up that (i) is prior to the patient's experience of authoring her action and (ii) is generally predictable of the patient's future decision. From (i) and (ii) scientists infer that this prior predicative neural buildup is actually the cause of the action, and not the agent's conscious intention. Al Mele (2014) has written a very helpful book detailing these arguments and the objections to these sorts of arguments. Among them, some criticize that these experiments are not testing the right sorts of actions (e.g., arbitrarily hitting a button is not a decision we would paradigmatically think of as freely willed), that these people in these experiments are incorrect about their own timing of authorship, or that this prior neural-build up is a kind of preparation event.

This is a kind of dialectical benefit for dualism because metaphysical theories of free will do acknowledge that (phenomenal) consciousness is a necessary condition for free will, but most theories do not adequately explain *how* consciousness is relevant for free will. As Timothy O'Connor (2000, 122) notes, the philosopher has an obligation to explain the relationship between consciousness and free will. He notes

It is a remarkable feature of most accounts of free will that they give no essential role to conscious awareness. One has the impression that an intelligent automata could conceivably satisfy the conditions set by these accounts—something that is very counterintuitive.

Gregg Caruso (2012, 143), agreeing with this passage from O'Connor, argues that according to folk psychology, it is simply “inconceivable” to imagine some sort of robot (assuming it lacks any sort of phenomenal consciousness) as a kind of free creature. He explains that robots (of this sort) may be able to do all sorts of functions, but when we think about free will, “we typically mean (among other things) that it was the result of a voluntarily choice, *consciously willed*” (ibid). This is why cases like epiphenomenalism, hypnosis, or overt manipulation (e.g., mind-control) are cases where we intuitively assume agents (subject to thesis conditions) are not free.

In explaining how phenomenal consciousness can be involved in microphysical states, the dualist needs to engage the Argument from Conservation Laws. However, if the dualist can show a plausible model of dualistic mental causation (as I discussed above) in the face of our current physics, then, arguably, a good share of the motivation to reject dualistic free will is left wanting, and the conjunction of dualism and free will is left as a plausible option.

6.3 Free Will Impossibilism

In this section, I shall discuss the second of the three views one might adopt as a reaction to the B-MODAL CONSEQUENCE ARGUMENT. This second option is what I call *free will impossibilism*.¹¹⁴ This is the view one would be attracted to if, for example, one (I) accepted the conclusion of my argument – that necessarily, if non-reductive physicalism is true, we do not have a free will - and (II) believed that non-reductive physicalism was true. Free will impossibilism is, thus, the view stating that no actual person can have a free will, even if they can still perhaps have rational, meaningful, and morally significant lives in other respects. This view goes back to at least Baruch Spinozza (1677), and has contemporary adherents in Galen Strawson (1962), Derk Pereboom (1995; 2001; 2011; 2014), Neil Levy (2011), and Greg Caruso (2012).

Notwithstanding the possible first reaction of some, free will impossibilism, at least in the vein of Pereboom (2014) and Caruso (2012), is not a nihilistic theory about norms and ethics.¹¹⁵ However, it is true that a common objection to Pereboom and Caruso is that their view is not livable (Vicens, 2015). Here, I am going to discuss some key aspects of the view, and show how the impossibilists can still affirm morals, values, rationality, and the meaning of life, which makes the view a viable option.

¹¹⁴ This is often referred to as free will skepticism or hard determinism. I prefer free will Impossibilism because the view is that free will is impossible in the actual world. Free will Impossibilism does not necessarily hold that free will is incoherent or metaphysically impossible, though, which may be a possible confusion with the title. However, I find both other titles slightly odd as well. Free will skepticism may give the impression that one is agnostic to whether there is free will in the actual world or not. Hard determinism sounds like the position is tied to the truth of determinism, which it is not.

¹¹⁵ See also Pereboom and Caruso (2022).

6.3.1 Moral Responsibility

Traditionally, we think that free will is at least a necessary condition for moral responsibility.¹¹⁶ So, if we deny human beings can have free will in the actual world, the obvious worry is that morality must be rejected wholesale. And Pereboom, partially acknowledges this, admitting that impossibilism “requires giving up our ordinary view of ourselves as blameworthy for immoral actions and praiseworthy for those that are morally exemplary” (Pereboom, 2011, p. 417). However, Pereboom also points out that rejecting the notion of blameworthiness and praiseworthiness does not entail that we have to give up our judgements that some actions are morally wrong. This is because, in fact, there are other aspects of morality besides the traditional concept of blameworthiness and praiseworthiness. Pereboom gives the following example: “[e]ven if we came to believe that some perpetrator of genocide was not blameworthy in this sense due to a degenerative brain disease, we would still hold that his actions were morally wrong, and that it was very bad that he acted as he did” (ibid). So, denying free will does not undercut our entire concept of morality, although it will require us to reject (or deeply revise) our notion of responsibility.

So, what would happen if everyone rejected the traditional notion of praiseworthiness and blameworthiness? Of course, in practice we assume that people are praiseworthy and blameworthy for their actions in both our practices of moral education and criminal justice system, and so these sorts of practices would have to be revised (ibid). This does not, however, mean that we have to

¹¹⁶ See fn. 10.

reject all criminal justice and educational practices. Pereboom points out that one compatible criminal justice theory with free will impossibilism is one that treats criminals analogous to the way societies treat carriers of infectious diseases (Pereboom, 2011, p. 418). A society may have a right to quarantine carriers of certain diseases, even if the people with these diseases did nothing morally blameworthy, but it is good for the society to place them in quarantine for the good of the future of the society.¹¹⁷ Analogously, we may not blame criminals, but we could still put them in prison, or other similar facilities, for the future good of the society. This might not be compatible with all of our contemporary criminal justice norms (death penalty, solitary confinement, no-parole practices), but it does show a possible way to justice at least some normal criminal justice practices.

6.3.2 Meaning of Life

And, if we think about the meaning of life, free will impossibilism may also seem to threaten our normal conception of life's meaning. What is the meaning of life? However, we can think about pretty intuitive test cases. Take a paradigm case of a meaningless life from Susan Wolf (2014, p. 92):

For me, the idea of a meaningless life is most clearly and effectively embodied in the image of a person who spends day after day, or night after night, in front of a television set, drinking beer and watching situation comedies. Not that I have anything against television or beer. Still the image, understood as an image of a person whose life is lived in hazy

¹¹⁷ See also Peter Strawson's classic paper Freedom and Resentment (Strawson, 1962).

passivity, a life lived at a not unpleasant level of consciousness, but unconnected to anyone or anything, going nowhere, achieving nothing—is, I submit, as strong an image of a meaningless life as there can be. Call this case The Blob.

Wolf (ibid) contrast “The Blob” with exemplars of meaningful lives:

If we step back, however, and ask ourselves, as observers, what lives strike us as especially meaningful, if we ask what sorts of lives exemplify meaningfulness, subjective criteria do not seem to be in the forefront. Who comes to mind? Perhaps Gandhi or Albert Schweitzer or Mother Teresa; perhaps Einstein or Jonas Salk. Cezanne or Manet, Beethoven or Charlie Parker...Alternatively, we can look to our neighbors, our colleagues, our relatives some of whom, it seems to me, live more meaningful lives than others. Some, indeed, of my acquaintance seem to me to live lives that are paradigms of meaning right up there with the famous names on the earlier lists; while others (perhaps despite their modicum of fame) would score quite low on the meaningfulness scale. If those in the latter category feel a lack of meaning in their lives—well, they are right to feel it, and it is a step in the right direction that they notice that there is something about their lives that they should try to change.

On Wolf’s view the difference between “The Blob” and these exemplary figures, such as Mother Teresa, is that these exemplars are “actively and at least somewhat successful engaged in projects (understanding this term broadly) that not just seem to have positive value, but that really do have it” (Wolf, 2014, p. 96). By project, she means some worthwhile endeavor. In these projects, the positive value is not just subjective meaning. Free will impossibilism would not entail that there are no values, or objectively good pursuits in their own right. Impossibilism would also not make planning of one’s efforts toward those goals impossible. So, it would not entail a wholesale

rejection of this sort of traditional view of the meaning of life. You could imagine a person, wholly accepting free will impossibilism, and yet striving for a meaningful life.

6.3.2 Rationality

The most unintuitive feature of the possibility of striving for a meaningful life, and accepting impossibilism, is imagining our decision making process in such endeavors. Imagine the Blob character from Wolf's example earlier contemplating on whether or not he should change his ways and get a job or join a charitable cause, and assume he fully believes in impossibilism. In his belief of impossibilism, he will accept that all of his decisions and actions are determined by causal antecedents beyond his control. He will also accept that he could not actually do anything differently than what he actual does, and yet he will still have emotional states like regret, shame, and hope driving him (Pereboom, 2014, p. 106). These sorts of reactive attitudes (e.g., guilt, shame, disgust) are usually motivational factors for moral- and life-transformations. The Blob might say to himself something like, "I should never have binged that Netflix show, I should have gotten involved in that Christmas charity program. I will never do that again!" But, with the knowledge of impossibilism, he would know that he could not have done anything besides what he did in the circumstances he was in. This would seem to at least lesson the possible motivation of his regret or guilt. While the impossibilists can accept a theory stating that one can be rational in deliberating *even* when one knows that one can only ever do the thing one actually does, it makes certain normal moral epistemological practices somewhat counterintuitive.

However, Caruso (2012) and Pereboom (2013) have gone to great pains to show that this view is *not* the nihilistic outcome that some believe it would be, and if one feels the argument for non-reductive physicalism is stronger than the argument for free will, then the B-MODAL CONSEQUENCE ARGUMENT would naturally lead one to this position.

6.4 Revisionism

The third of three possible views one could accept on the basis of the B-MODAL CONSEQUENCE ARGUMENT is called *revisionism*. Revisionism is a hybrid view. It is both a kind of possibilist view and an innovative form of compatibilism (Pereboom & McKenna, 2016, p. 286). Kelly McCormick characterizes revisionism as “the view that we can and should distinguish between what we think about [free will] and what we ought to think about [free will], that the former is in some important sense implausible and conflicts with the latter, and so we should revise our concept accordingly” (McCormick, 2017, p. 109). This is the view one could be attracted to if, for example, one (I) accepted the conclusion of my argument – that necessarily if non-reductive physicalism is true, then no one has free will– and (II) believed non-reductive physicalism was true, but (III) found free will impossibilism unappealing for some reason. The revisionist view is held by several philosophers including prominently Manuel Vargas (2005; 2009; 2011; 2014), Eddy Nahmias (2014), Kelly McCormick (2013; 2014; 2015; 2017), Stephen Nichols (2007), and, more recently, Oisín Deery (2020).¹¹⁸

Revisionism is a nontraditional view (or, simply, a more recent family of views) for understanding free will. Kelly McCormick explains that the central difference between Revisionism and other, more traditional, views of free will, is the revisionist understanding of the distinction between *diagnostic* and *prescriptive* approaches to free will and moral responsibility (2017, p. 109). A diagnostic account of free will and responsibility is one that “most accurately reflects our widely

¹¹⁸ Deery’s view (2020) is quite interesting in that he views free will as a natural kind. That is a new take on revisionism. I am not going to discuss his view in depth here, but it is worth further consideration.

shared commitments as they stand” (ibid). In other words, these analyses *describe* what we do take the common concept or folk idea of free will. By contrast, a prescriptive approach provides an account of what we *ought* to take free will to be, all things considered (ibid).¹¹⁹ Revisionism is a prescriptive approach to free will.

Because revisionism is a prescriptive approach, revisionists typically agree with the incompatibilist philosophers that our intuitive or folk picture of free will is an incompatibilist one (e.g., requiring alternative possibilities in some deep sense). McCormick, however, continues (2017, p. 110) by claiming that “revisionists will identify some feature of our [incompatibilist] folk concept that is deeply problematic, and argue that we ought to abandon this feature and revise our concept accordingly.” But, like other compatibilists, revisionists do believe that we have free will in the actual world, when we consider the new, revised concept of free will.

Here, Manuel Vargas (2011, pg., 458-459) offers to explain the view: imagine a theist named Friedrich who believes earnestly that (i) there are genuine necessary moral truths (e.g., torturing babies for fun is wrong, giving alms to the poor is good) and (ii) moral truths are, in some sense, determined by God (e.g., perhaps they are grounded in God’s nature). So, if you asked Friedrich what is right and wrong, he might say something like “whatever God says is right and wrong.” But, Vargas adds that we are to imagine what would happen to Friedrich if, for whatever reason, he became an atheist. In that circumstance, what would Friedrich do about his beliefs regarding morality? One possibility is that, given (i) and (ii), Friedrich might “conclude that morality is a

¹¹⁹ And, these view typically has some empirical or naturalistic criteria for what we need to take into account (Vargas, 2011, pp. 59-60).

sham” (ibid., 458). So both (i) and (ii) are rejected. Friedrich might live his life *as if morality existed* as a pragmatic matter (although he might not), but what he would do for sure is eliminate moral truths from his list of facts about the world (at least as a metaphysical matter). This is what we can call the eliminativist strategy. However, Friedrich might opt for a different strategy, he might instead *only* reject (ii). Vargas notes, “This might be an especially tempting option if he came to believe that morality’s distinctive judgments and practices could be grounded in practical reason (or sentiment, or human functioning, or ideal social arrangements with which no one could reasonably disagree, etc.)” (ibid, 459). Of course, Friedrich would have to acknowledge that his beliefs about morality have changed, he used to believe in (i) and (ii)-that moral truths were real and that God determined them, but he now only believes in (i). “Nevertheless, if he reasoned in this way he would likely insist that morality should be included in the catalog of what truly exists in our world” (ibid, pg., 458-459). This is the revisionist option, which can be applied to free will (and other metaphysical views).

In this analogy, Frederick’s eliminativism option about morality is analogous to free will impossibilism. Back in Chapter One (§1.1), I defined free will as having two key conditions – CONSCIOUS QUAUSATION and ALTERNATIVE POSSIBILITIES – that I claimed were intuitive and prominent in folk psychology. The revisionist idea would be that “abandon some commitments [like ALTERNATIVE POSSIBILITIES] that constitute our ordinary way of thinking about free will” and, once we have done that, we can acknowledge that people actually do have free will in the actual world (Vargas, 2011, pg., 460). And, some philosophers, particularly inspired by Frankfurt-styled cases (Sartorio, 2016), prefer an account of free will without ALTERNATIVE POSSIBILITIES, usually called *actual-sequence freedom*. So, one could use the B-MODAL CONSEQUENCE ARGUMENT in conjunction with non-reductive physicalism to argue for a kind of actual-sequence view of freedom.

I think methodologically, there is going to be a question over when revisionism is the appropriate approach about our concepts (as opposed to simply eliminating them). Some philosophers will find the impossibilist's answer to be simpler than the revisionists, but compatibilists may find this a useful strategy to respond to the B-MODAL CONSEQUENCE ARGUMENT.

Conclusion

Interactionist dualism, impossibilism, and revisionism are all possible reactions to the consequence argument's success. If the sort of mind-body interactionism described in §6.2 is not plausible, then one will naturally opt either for impossibilism or revisionism. As it pertains to these two views, the main question here is methodological: When is it acceptable to revise our metaphysical concepts and when should we eliminate them? Philosophers will answer this two-part question differently, but this question will largely determine whether impossibilism or revisionism is the preferable option. For my own part, it seems intuitive that moving to impossibilism or revisionism would be less preferable reactions, all things being equal, than a kind of interactionist dualism. This preference is mostly based on the fact that among these three options, interactionist dualism is the only one that fully preserves the concept of free will. However, the question of which of these three views is ultimately right is one I will leave for further research. I started my thesis with the question whether or not God needs to reach up to the second shelf (i.e., the non-physical shelf) in order to create free creatures. The B-Modal Consequence Argument provides a reason to believe that God would indeed have to reach up to that second shelf in order to create free creatures.

REFERENCES

- Albert, D., 1992. *Quantum Mechanics and Experience*. Cambridge (MA): Harvard University Press.
- Albert, D., 2016. Physics and Chance. In: *After Physics*. Cambridge (MA): Harvard University Press.
- Armstrong, D., 1983. *What is a Law of Nature?*. Cambridge (UK): Cambridge University Press.
- Bailey, A. & Rasmussen, J., 2020. A new puppet puzzle. *Philosophical Explorations*, 23(3), pp. 202-213.
- Baker, L. R., 2009. *The Metaphysics of Everyday Life: An Essay in Practical Realism (Cambridge Studies in Philosophy)*. Cambridge, MA: Cambridge University Press.
- Barrett, J., 2019. *The Conceptual Foundations of Quantum Mechanics*. New York: Oxford University Press.
- Baysan, U., 2019. Emergence, Function, and Realization. In: S. G. e. a. (eds.), ed. *Routledge Handbook of Emergence*. New York: Routledge.
- Beebe, H., 2000. The Non-Governing Conception of Laws of Nature. *Philosophy and Phenomenological Research*, 61(3), pp. 571-594.
- Bennett, K., 2007. Mental Causation. *Philosophy Compass*, 2(2), pp. 316-337.
- Bernstein, S. & Wilson, J. M., 2016. Free will and mental quausation. *Journal of the American Philosophical Association*, 2(2), pp. 310-331.
- Blackmore, S., 2006. *Conversations on Consciousness : What the Best Minds Think about the Brain, Free Will, and What It Means to Be Human*. New York: Oxford University Press.
- Block, N., 2003. Do Causal Powers Drain Away?. *Philsoophy and Phenomenological Research* , Volume 67, pp. 133-150.
- Brateman, M., 1987. *Intention, Plans, and Practical Reason*. Cambridge (MA): Harvard University Press.
- Bunge, M., 1980. *The mind-body problem: A psychobiological approach..* Oxford (UK): Pergamon.

- Campbell, J., 2007. Free Will and the Necessity of the Past. *Analysis*, Volume 67, pp. 105-111.
- Campbell, J., 2010. Incompatabilism and Faralism: Reply to Loss. *Analysis*, Volume 70, pp. 71-76.
- Campbell, J., 2017. The Consequence Argument. In: *The Routledge Companion on Free Will*. New York: Routledge, pp. 151-166.
- Carroll, J. & Markosian, N., 2010. *An Introduction to Metaphysics*. New York: Cambridge University Press.
- Cartwright, N., 2022. *A Philosopher Looks At Science*. New York: Cambridge University Press.
- Caruso, G., 2012. *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Plymouth (UK): Lexington Books.
- Chalmers, D., 1996. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
- Chalmers, D., 2002. Does conceivability entail possibility. In: T. S. Gendler & J. Hawthorne, eds. *Conceivability and Possibility*. New York: Oxford University Press, pp. 145--200.
- Chalmers, D., 2010. *The Character of Consciousness*. New York: Oxford University Press.
- Clark, S. & Thomas, I., 2019. Emergence in Non-Relativistic Quantum Mechanics. In: S. Gibb, R. Hendry & T. Lancaster, eds. *The Routldge Handbook on Emergence*. New York: Routledge, pp. 265-275.
- Collins, R., 2008. Modern Physics and the Energy-Conservation Objection to Mind-Body Dualism. *American Philosophical Quarterly*, 45(1), pp. 31-42.
- Cover, J. & O'Leary-Hawthorne, J., 1996. Free Agency and Materialism. In: J. Jordan & D. Howard-Synder, eds. *Faith Freedom and Rationality*. Lanham (MD): Rowman & Littlefield, pp. 47-72.
- Cyr, T., 2017. Is Semicompatibilism Unstable?. *Disputatio*, 9(45), pp. 245-264.
- Cyr, T., 2022. The Robustness Requirement on Alternative Possibilities. *The Journal of Ethics*, 26(3), pp. 481-499.
- Davidson, D., 1963. Actions, Reasons, and Causes. *The Journal of Philosophy*, 60(23), pp. 685-700.
- Deery, O., 2021. *Naturally Free Action*. New York: Oxford University Press.
- Deery, O., Bedke, m. & Nichols, S., 2013. Phenomenal Abilities: Incompatibilism and the Experience of Agency. In: D. Shoemaker, ed. *Oxford Studies in Agency and Responsibility (Vol 1)*. New York: Oxford University Press, pp. 126-150.
- Dennett, D., 1991. *Concioussness Explained*. Boston: Little, Brown and Company.

- Dennett, D., 2015. *Elbow Room: Varieties of Free Will Worth Wanting*. 2 ed. Cambridge (MA): Bradford Books.
- Dimitrijevic, D., 2020. Causal closure of the physical, mental. *European Journal for Philosophy of Science* , 10(1), pp. 1-22.
- Dorr, C., 2016. Agaisnt Counterfactual Miracles. *Philosophical Review*, Volume 125, p. 241–86.
- Dorr, C., 2019. Natural Properties. *Stanford Encyclopedia of Philosophy*.
- Dowell, J., 2006. The physical: Empirical, not metaphysical. *Philosophical Studies*, Issue 131, pp. 25-60.
- Dretske, F., 1977. Laws of nature. *Philosophy of Science* , 44(2), pp. 248-268.
- Earman, J., 1986. *A Primer on Determinism (The Western Ontario Series in Philosophy of Science, 32)*. New York: Springer.
- Earman, J. & Norton, J., 1999. EXORCIST XIV: The Wrath of Maxwell's Demon. Part II. From Szilard to Landauer and Beyond. *STUDIES IN HISTORY AND PHILOSOPHY OF MODERN PHYSICS*, 30(1), pp. 1-30.
- Finch, A. & Warfield, T., 1998. The MIND Argument and Libertarianism. *Mind*, Volume 127, pp. 515-528.
- Fischer, J. M., 1994. *The metaphysics of free will: An essay on control..* New York: Oxford University Press.
- Fischer, J. M., 2022. The Frankfurt-style cases: extinguishing the flickers. *Inquiry: An International Journal of Philosophy*, 65(9), pp. 1185-1209.
- Fodor, J., 1987. *Psychosemantics*. Cambridge (MA): MIT Press.
- Fodor, J., 1994. "The Mind-Body Problem," in The Mind-Body Problem, ed. R. Warner and T.. In: R. W. a. T. Suzba, ed. *The Mind-Body Problem*. Oxford: Blackwell.
- Franklin, C., 2016. Review of 'Causation and Free Will'. *Notre Dame Philosophical Reviews*.
- Furlong, P., 2019. *The Challenges of Divine Determinism: A Philosophical Analysis*. New York: Cambridge University Press.
- Gibb, S., 2010. Closure Principles and the Laws of Conservation of Energy and Momentum. *Dialectica*, 64(3), pp. 363-384.
- Gibb, S., 2015. The Causal Closure Principle. *The Philosophical Quarterly*, 65(261), p. 626–647.
- Goetz, S., 2005. Frankfurt-Style Counterexamples and Begging the Question. *Midwest Studies*, Volume 25, p. 83–105.
- Goff, P., 2017. *Consciousness and Fundamental Reality*. New York: Oxford University Press.

- Goldman, N., 1954. *Fact, Fiction and Forecast*. 4th ed. Cambridge (MA): Harvard University Press.
- Haggard, P. & Libet, B., 2001. Conscious intention and brain activity. *Journal of Consciousness Studies*, 8 (11), pp. 47-63.
- Hájek, A., 2021. Contra counterfactualism. *Synthese*, Volume 199, p. 181–210.
- Heil, J., 2020. *Philosophy of Mind: A Contemporary Introduction*. New York: Routledge.
- Hempel, C., 1980. Comments on Goodman's Ways of Worldmaking. *Synthese*, pp. 139-199.
- Hieronimi, P., 2022. What is a will?. In: U. Maoz & W. Sinnott-Armstrong, eds. *Free Will: Philosophers and Neuroscientists*. New York: Oxford University Press, pp. 13-20.
- Hildebrand, T., 2023. *Laws of Nature*. New York: Cambridge University Press.
- Hitchcock, C., 2018. Probabilistic Causation. *Stanford Encyclopedia of Philosophy*.
- Horgan, T., 1989. Mental Quausation. *Philosophical Perspectives*, Volume 3, pp. 47-76.
- Horgan, T., Maslen, C. & Daly, H., 2009. Mental Causation. In: H. Beebe, C. Hitchcock & P. Menzies, eds. *The Oxford Handbook of Causation*. New York: Oxford University Press.
- Huemer, M., 2000. Van Inwagen's Consequence Argument. *Philosophical Review*, Volume 109, pp. 524-43.
- Huemer, M., 2004. Elusive Freedom? A Reply to Helen Beebe. *Philosophical Review*, Volume 113, pp. 411-416.
- Huemer, M., 2021. *Knowledge, Reality, and Value*. s.l.:s.n.
- Jackson, F., 1998. *From Metaphysics To Ethics: A Defense of Conceptual Analysis*. New York: Oxford University Press.
- Kane, R., 1996. *The Significance of Free Will*. New York: Oxford University Press.
- Kant, I., 1785. *Fundamental Principles of the Metaphysic of Morals*. Ebook: GUTENBERG.
- Kim, J., 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation (Representation and Mind series)*. New York: Bradford Book.
- Kim, J., 2005. *Physicalism, or Something Near Enough (Princeton Monographs in Philosophy, 19)*. Princeton, NJ : Princeton University Press.
- Kim, J., 2010. *Essays in the Metaphysics of Mind*. New York: Oxford University Press.
- Kocurek, A. W., 2022. Does Chance Undermine Would?. *Mind*, 131 (523), pp. 747-787.
- Korman, D., 2020. Ordinary Objects. *Stanford Encyclopedia of Philosophy*.

- Kratzer, A., 1977. What 'Must' and 'Can' Must and Can Mean. *Linguistics and Philosophy*, Volume 1, p. 337–355.
- Kripke, S., 1980. *Naming and Necessity*. Cambridge: Cambridge University Press.
- Lambert, F. & Merluzzi, P., 2021. Counterfactuals, counteractuals, and Free Choice. *Philosophical Studies*, 178(2), p. 445–469.
- Lewis, D., 1966. An Argument For The Identity Theory. *Journal of Philosophy*, Volume 63, pp. 17-25.
- Lewis, D., 1973. *Counterfactuals*. New York: Oxford:Blackwell.
- Lewis, D., 1979. Counterfactual Dependence and Time's Arrow. *Nous*, 13(4), pp. 455-476.
- Lewis, D., 1981. Are we free to break the laws?. *Theoria*, 47(3), pp. 113-121.
- Lewis, D., 1986. *On the Plurality of Worlds*. New York: Oxford University Press.
- Lewis, D., 1994. Reduction of Mind. In: S. Guttenplan, ed. *A Companion to Philosophy of Mind*. New York: Oxford: Blackwell Publishers, pp. 412-431.
- Lewis, D., 2002. Tensing the Copula. *Mind*, Volume 111, pp. 1-14.
- Lewis, D., 2004. Void and Object. In: L. Paul & N. Collins, eds. *Causation and Counterfactuals*. Cambridge: MIT Press, p. 77–290.
- Lewis, D. & Langton, R., 1998. Defining 'Intrinsic'. *Philosophy and Phenomenological Research*, Volume 58, p. 335.
- Libet, B., 1985. Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. *Behavioral and Brain Sciences*, Volume 8, pp. 529-566.
- Libet, B., 2004. *Mind Time*. Cambridge, MA: Harvard University Press.
- List, C., 2019. What's Wrong with the Consequence Argument: A Compatibilist Libertarian Response. *Proceedings of the Aristotelian Society*, 119(3), pp. 253-274.
- List, C., 2019. *Why Free Will Is Real*. Cambridge, MA: Harvard University Press.
- List, C. & Menzies, P., 2009. Non-Reductive Physicalism and the Limits of the Exclusion Principle. *Journal of Philosophy*, Volume 9, pp. 475-502.
- Loar, B., 2017. *Consciousness and Meaning: Selected Essays*. New York: Oxford University Press.
- Loar, B., 2017. Phenomenal States. In: K. Balog & S. Beardman, eds. *Consciousness and Meaning: Selected Essays*. Oxford: Oxford University Press.

- Loewer, B., 2001. From Physics to Physicalism. In: C. Gillett & B. Loewer, eds. *Physicalism and its Discontents*. New York: Cambridge University Press, pp. 37-56.
- Loewer, B., 2009. Why Is There Anything Except Physics?. *Synthese*, 170(2), pp. 217-233.
- Loewer, B., 2015. Mental Causation: The Free Lunch. In: T. Horgan, M. Sabates & D. Sosa, eds. *Qualia and Mental Causation in a Physical World: Themes from the Philosophy of Jaegwon Kim*. Cambridge (MA): Cambridge University Press, pp. 40-64.
- Loewer, B., Forthcoming. "The Consequence Argument Meets the Mentaculus. In: Loewer, Weslake & Winberb, eds. *Time's Arrow and the Probability Structure of the World*. Boston, MA: Harvard University Press.
- Lowe, E., 2000. Causal Closure Principles and Emergentism. *Philosophy (Cambridge University Press)*, 75(4), pp. 571 - 585.
- Lowe, E., 2005. *The Four-Category Ontology: A Metaphysical Foundation for Natural Science*. Oxford: Oxford University Press.
- Mackie, P., 2015. Fatalism, Incompatabilism, and The Ability To Do Otherwise. In: J. M. F. a. P. Todd, ed. *Freedom, Fatalism, and Foreknowledge*. New York: Oxford University Press, pp. 128-146.
- Maier, J., 2015. The Agentive Modalities. *Philosophy and Phenomenological Research*, XC(1), pp. 113-134.
- Maier, J., 2020. Abilities. *Stanford Encyclopedia of Philosophy*.
- Maudlin, T., 2007. *The Metaphysics Within Physics*. New York: Oxford University Press.
- Maxwell, J. C., 2011. *Theory of Heat*. Cambridge: Cambridge University Press.
- McCormick, K., 2017. Revisionism. In: K. Timpe, M. Griffith & N. Levy, eds. *Routledge Companion to Free Will*. New York: Routledge, pp. 109-120.
- McKay, T. & Johnson, D., 1996. A Reconsideration of an Argument Against Compatibilism. *Philosophical Topics*, Volume 24, pp. 113-122.
- McLaughlin, B., 2013. Why Rationalization Is Not a Species of Causal Explanation. In: J. D'Oro, ed. *Reasons and Causes: Causalism and anti-Causalism in the Philosophy of Action*. London: Palgrave MacMillan.
- Mele, A., 2006. *Free Will and Luck*. New York: Oxford University Press.
- Mele, A., 2009. *Effective Intentions*. New York: Oxford University Press.
- Mele, A., 2014. *Free: Why Science Hasn't Disproved Free Will*. New York: Oxford University Press.
- Mele, A., 2022. *Free Will: An Opinionated Guide*. New York: Oxford University Press.

- Mele, A. & Beebe, H., 2002. Humean Compatibilism. *Mind*, 111(442), pp. 201-223 .
- Mele, A. & Robb, D., 1998. Rescuing Frankfurt-Style Cases. *Philosophical Review*, 107(1), pp. 97-112.
- Mele, A. & Robb, D., 2003. Bbs, magnets and seesaws: The metaphysics of Frankfurtstyle. In: M. & W. D. McKenna, ed. *Moral Responsibility and Alternative Possibilities*. Aldershot: Ashgate.
- Melnyk, A., 2007. *A Physicalist Manifesto: Thoroughly Modern Materialism*. New York: Cambridge.
- Mendelovici, A., 2018. *Phenomenal Basis of Intentionality*. New York: Oxford University Press.
- Menzies, P. & Beebe, H., 2020. Counterfactual Theories of Causation. *The Stanford Encyclopedia of Philosophy*, Volume Winter 2020 Edition.
- Mumford, S., 2004. *Laws in Nature (Routledge Studies in Twentieth-Century Philosophy)*. New York: Routledge.
- Mumford, S., 2010. Causal Powers and Capacities. In: H. Beebe, C. Hitchcock & P. Menzies, eds. *Oxford Handbook on Causation*. Oxford: Oxford University Press, p. 265–278.
- Nahmias, E., 2014. Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences. In: W. Sinnott-Armstrong, ed. *Moral Psychology, vol. 4: Freedom and Responsibility*. Cambridge (MA): MIT Press.
- Ney, A., 2008. Physicalism as an attitude. *Philosophical studies*, 138(1), pp. 1-15.
- Ney, A., 2020. Are the Questions of Metaphysics More. *Philosophy and Phenomenological Research*, 100(3), pp. 696-715.
- Ney, A., 2020. The Fundamentality of Physics: Completeness or Maximality?. In: K. Bennett & D. Zimmerman, eds. *Oxford Studies in Metaphysics Volume 12*. New York: Oxford University Press, pp. 204-228.
- Ney, A., 2021. *The World in the Wave Function: A Metaphysics for Quantum Physics*. New York: Oxford University Press.
- Norton, J., 2013. *John D. Norton Personal Website*. [Online] Available at: https://sites.pitt.edu/~jdnorton/Goodies/exorcism_phase_vol/exorcism_phase_vol.html [Accessed 4th June 2022].
- O'brien, L., 2014. *Philosophy of Action (Palgrave Philosophy Today)*. New York: Palgrave Macmillan.

- O'Connor, T., 2021. Free Will in a Network of Interacting Causes. In: W. Simpson, R. Koons & J. Orr, eds. *Neo-Aristotelian Metaphysics and the Theology of Nature*. New York: Routledge, pp. 151-169.
- Oderberg, D., 2016. Divine Premotion. *International Journal for Philosophy of Religion*, 79(3), pp. 207-222.
- Palmer, D. & Ginet, C., 2010. On Mele and Robb's Indeterminate Frankfurt-styled Case. *Philosophy and Phenomenological Research*, LXXX(2), pp. 440-446.
- Papineau, D., 1998. Mind the Gap. *Philosophical Perspectives*, 32(12), pp. 373-388.
- Papineau, D., 2001. The Rise of Physicalism. In: C. Gillett & B. Loewer, eds. *Physicalism and its Discontents*. Cambridge: Cambridge University Press, pp. 3-36.
- Papineau, D., 2002. *Thinking About Consciousness*. New York: Oxford University Press.
- Papineau, D., 2008. Must a physicalist be a microphysicalist?. In: J. H. & J. Kallestrup, ed. *Being Reduced: New Essays on Reduction, Explanation, and Causation*. New York: Oxford University Press, pp. 126-148.
- Papineau, D., 2009. The Causal Closure of the Physical and Naturalism. In: A. Beckermann, B. P. McLaughlin & S. Walter, eds. *The Oxford Handbook of Philosophy of Mind*. New York: Oxford University Press, pp. 53-65.
- Papineau, D., 2013. Causation is Macroscopic but Not Irreducible. In: *Mental Causation and Ontology*. New York: Oxford University Press, p. 126-152.
- Paul, L., 2007. Constitutive Overdeterminism. *Topics in Contemporary Philosophy*.
- Pearce, K., 2022. Foundational Grounding and Creaturely Freedom. *Mind*, 131(524), p. 1108-1130.
- Pereboom, D., 2011. *Consciousness and the prospects of physicalism*. New York: Oxford University Press.
- Pereboom, D., 2011. Free-Will Skepticism and the Meaning in Life. In: R. Kane, ed. *Oxford Handbook of Free Will*. New York: Oxford University Press, pp. 407-424.
- Pereboom, D., 2014. *Free Will, Agency, and Meaning in Life*. New York: Oxford.
- Pereboom, D., 2022. *Free Will: Elements in Philosophy of Mind*. Cambridge: Cambridge University Press.
- Pereboom, D. & McKenna, M., 2016. *Free Will: A Contemporary Introduction (Routledge Contemporary Introductions to Philosophy)*. New York: Routledge.
- Pietroski, P., 2000 . *Causing Actions*. New York: Oxford University Press.

- Pitts, J. B., 2021. Conservation of Energy: Missing Features in Its Nature. *Foundations of Science*, Volume 26, p. 559–584.
- Plantinga, A., 1967. *God and other minds: a study of the rational justification of belief in God*. Ithaca, NY: Cornell University Press.
- Plantinga, A., 2011. *Where the Conflict Really Lies: Science, Religion, and Naturalism*. New York: Oxford University Press.
- Pruss, A., 2013. Incompatibilism Proved. *Canadian Journal of Philosophy*, 43(4), pp. 430 - 437.
- Rasmussen, J., 2023. *Who Are You, Really?: A Philosopher's Inquiry into the Nature and Origin of Persons*. Downers Grove (IL): IVP Academic.
- Robb, D., 2020. Moral Responsibility and the Principle of Alternative Possibilities. *The Stanford Encyclopedia of Philosophy (Fall 2020 Edition)*.
- Robb, D. & Mele, A., 1998. Rescuing Frankfurt-style cases. *Philosophical Review*, 107(1), pp. 97-112.
- Sartorio, C., 2016. *Causation and Free Will*. New York: Oxford University Press.
- Sartorio, C. & Kane, R., 2022. *Do We Have Free Will? (Little Debates about Big Questions)*. New York: Routledge.
- Schaffer, J., 2009. Spacetime the one substance. *Philosophical Studies*, Volume 145, p. 131–148.
- Schlegel, A. et al., 2015. Hypnotizing Libet: Readiness potentials with non-conscious. *Consciousness and Cognition*, Volume 33, pp. 196-203.
- Schwartz, J. H. S. a. M. B., 2003. Quantum Physics in Neuroscience and. *Philosophical Transactions of the Royal Library*.
- Searle, J., 2004. *Mind: A Brief Introduction*. New York: Oxford University Press.
- Sider, T., 2003. What's So Bad about Overdetermination?. *Philosophy and Phenomenological Research*, Volume 67, p. 719–726.
- Sinnott-Armstrong, W., 2022. What is freedom?. In: U. Maoz & W. Sinnott-Armstrong, eds. *Free Will: Philosophers and Neuroscientists in Conversation*. New York: Oxford University Press, pp. 34-40.
- Slote, M., 1982. Selective Necessity and the Free-Will Problem. *Journal of Philosophy*, Volume 79, pp. 5-24.
- Smith, M., 2010. Practical Imagination and its Limits. *Philosophers'*, 10(3), pp. 1-20.
- Smith, P. & Jones, O., 1986. *Philosophy of Mind*. Cambridge: Cambridge University Press.

- Soon, C., Brass, M., Heinze, H.-J. & Haynes, J.-D., 2008. Unconscious Determinants of Free Decisions in the Human Brain. *Nature Neuroscience*, Volume 11, pp. 543-545.
- Speak, D., 2011. The Consequence Argument Revisited. In: R. Kane, ed. *The Oxford Handbook on Free Will*. New York: Oxford University Press, pp. 115-130.
- Spencer, J., 2017. Able to Do the Impossible. *Mind* , Volume 126, pp. 465-96.
- Stefansson, H. O., 2018. Counterfactual Skepticism and Multidimensional. *Erkenn* , Volume 83, pp. 875-898.
- Steward, H., 2012 . Actions as Processes. *Philosophical Perspectives*, Vol 26(Philosophy of Mind (2012)), pp. 373-388 .
- Steward, H., 2014. *A Metaphysics For Freedom*. New York: Oxford University Press.
- Steward, H., 2016. Libertarianism as a Naturalistic Position. In: K. Timpl & D. Speak, eds. *Free Will and Theism*. New York: Oxford University Press, pp. 158-171.
- Stoljar, D., 2022. Physicalism. In: E. N. Zalta, ed. *The Stanford Encyclopedia of Philosophy (Summer 2022 Edition)*. s.l.:s.n.
- Strawson, P., 1962. Freedom and Resentment. *Proceedings of the British Academy* , Volume 48, pp. 187-211.
- Swinburne, R., 2006. Relations Between Universals, or Divine Laws?. *Australasian Journal of Philosophy*, Volume 84, pp. 179-189.
- Taylor, C., 1964. *The Explanation of Behavior*. London: Routledge & Kegan Paul..
- Todd, P., 2021. *The Open Future: Why Future*. New York: Oxford University Press.
- Todd, P. & Fischer, J. M., 2015. Introduction. In: P. Todd & J. M. Fischer, eds. *Freedom, Fatalism, and Foreknowledge*. New York: Oxford University Press, pp. 1-40.
- Tooley, M., 1977. The Nature of Laws. *Canadian Journal of Philosophy*, 7(4), pp. 667-98 .
- Van Fraassen, B., 2002. *The Empirical Stance*. New Haven: Yale University Press.
- Van Inwagen, P., 1983. *An Essay on Free Will*. New York: Oxford University Press.
- Van Inwagen, P., 1989. When is the Will Free?. *Philosophical Perspectives*, Volume 3, pp. 399-422.
- Van Inwagen, P., 2008. Mystery of Metaphysical Freedom. In: P. Van Inwagen & D. Zimmerman, eds. *Metaphysics: the big questions*. Malden, MA: Blackwell, pp. 456-465.
- Van Inwagen, P., 2015. Some Thoughts on An Essay on Free Will. *The Harvard Review of Philosophy*, Volume XXII, pp. 16-30.

- Van Inwagen, P., 2017. *Thinking about Free Will*. New York: Cambridge University Press.
- Vargas, M., 2011. Revisionist Accounts of Free Will: Origins, Varieties, and Challenges. In: R. Kane, ed. *Oxford Handbook on Free Will*. 2 ed. New York: Oxford University Press.
- Vargas, M., 2011. Revisionist accounts of free will: Origins, varieties, and challenges. . In: R. Kane, ed. *The Oxford handbook of free will*. New York: Oxford University Press, p. 457–484.
- Vetter, B., 2018. *Potentiality: From Dispositions to Modality*. New York: Oxford University Press.
- Vicens, L., 2015. Derk Pereboom, FREE WILL, AGENCY, AND MEANING IN LIFE. *Faith and Philosophy: Journal of the Society of Christian Philosophers*, 32(2), pp. 230-235.
- Vihvelin, K., 2013. *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. New York: Oxford University Press.
- Vihvelin, K. & Tomkow, T., 2016. *Kadri Vihvelin's Philosophical Blog*. [Online] Available at: <http://vihvelin.com/> [Accessed 11 03 2023].
- Warfield, T., 2000. Causal Determinism and Human Freedom are Incompatible: A New Argument For Incompatibilism. *Philosophical Perspectives*, Volume 14, pp. 167-180.
- Weatherson, B., 2021. David Lewis. *Stanford Encyclopedia of Philosophy*.
- Wegner, D., 2002. *The Illusion of Conscious Will*. New York: Bradford Books.
- Wessling, J. & Turner, P. R., 2021. W. Matthews Grant on Human Free Will, and Divine Universal Causation. *Faith and Philosophy*, 38(3), pp. 313-336.
- Widerker, D., 1987. On an argument for Incompatibilism. *analysis*, 46(4), pp. 37-41.
- Widerker, D., 2019. Pereboom's Defense of Deliberation-Compatibilism: A Problem Remains. *The Journal of Ethics*, 23(3), pp. 333-345.
- Wiggins, D., 1968. On Being in the Same Place at the Same Time. *Philosophical Review*, Volume 77, pp. 90-95.
- Wiggins, D., 1980. *Sameness and Substance*. Oxford: Blackwell.
- Wigner, E., 1961. Remarks on the Mind-Body Problem. In: U. Good, ed. *The Scientist Speculates*. New York: Basic Books, pp. 284-302.
- Wilson, J., 2011. Non-reductive Realization and the Powers-based Subset Strategy. *The Monist*, 94(1), p. 121–154.
- Wilson, J., 2021. *Metaphysical Emergence*. New York: Oxford University Press.

Wolf, S., 2014. *The Variety of Values: Essays on Morality, Meaning, and Love*. New York: Oxford University Press.