What Would Have Been and What Should Have Been: The Interdependence of Causation and Morality

Laura Fearnley

Submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy

Philosophy
School of Humanities
College of Arts
University of Glasgow

November 2022

Abstract

This thesis is about morality, causation and the connection between the two. Whether there's some causal relation between flicking the switch and turning on the light, between donating blood and saving a life, or between rain falling and puddles on the ground, is typically understood to be a mind-independent, objective, precise matter of fact. It's no surprise given this perspective that for a long time philosophers didn't believe something so ostensibly nebulous as morality could be a determiner of causal relations. However, recent contributions to the literature have begun to pushback against this platitude by using moral considerations, alongside a whole host of other normative notions, to determine what causal facts there are in the world. One aim of this thesis is to contribute to this recent research by arguing that an appeal to moral considerations furnishes a causal account with the resources to meet various desiderata associated with theories of causation. Insofar as this is the case, I argue that causal accounts which incorporate moral considerations are more successful than those which do not. Thus, I argue that facts about causality partly depend upon facts about morality.

Causation's effect on morality is much less controversial. Moral philosophers agree that the truth of some moral claims partly depend upon what causal relations there are in the world. This is most noticeably true in the domain of moral responsibility. The thought is that if we are to be morally responsible for some event that is not itself an aspect for our conduct, then there must be some metaphysical relation — some metaphysical glue — that links our conduct to the event in question. As I and most others see it, there is one such relation: causation. Despite this consensus, there has been surprisingly little exploration into what kinds of causal facts determine the truth of moral assessments. I suspect this lacuna has arisen because it is tempting for those working in ethics to remain neutral on questions regarding causation, lest she have to face up to dealing with the thorny issues such questions invite. However, sometimes it is only by getting involved in such issues that progress can be made. With this in mind, the second aim of the thesis is to advance this debate by clarifying and evaluating what kind of causal facts determine moral facts. In particular, I address what kind of causal facts determine whether one is morally responsible for an outcome, and whether and to what extent one is praiseworthy for that outcome. With regards to assessments of moral responsibility, I argue that whether one can be held morally responsible for the effects of an omission depends upon the causal stability of that omission. In regards to praiseworthiness, I argue that whether and to what extent one is

praiseworthy for doing the right thing depends upon how causally robust one's motivation was for acting. Thus, I argue that facts about morality are partly determined by facts about causality.

In making these arguments, I defend the view that causal facts partly depend upon moral facts, and that moral facts partly depend upon causal facts. This lends support to a novel claim; namely, that morality and causation are interdependent.

Table of Contents

List of Tables and Figures

Acknowledgements

First and foremost, thank you to my supervisors, Robert Cowan and Neil McDonnell. It's hard to put into words just how much my academic time has been shaped by their influence. I will always be grateful for their time, patience, enthusiasm and steady encouragement. Their insightful and incisive guidance has made this thesis immeasurably better. I am also indebted to both of them for the many academic opportunities they've made available to me during the past four years, these opportunities have made the PhD considerably better than it otherwise would have been. I am very lucky to have had this supervisory team.

The wider philosophical community at Glasgow have been enormously supportive, particularly the denizens of postgraduate office 208 past and present. Thank you to Eilidh Harrison, María Pía Méndez Mateluna, Suzi Murning, Dario Mortini, Cody Cantabrana, Adriana Alcaraz-Sánchez, Finn McCardel, Martin Miragoli, Lou Logan James Humphries and Joe Slater for their warmth and friendship over the years, and for enabling my compulsion to go to the pub way before it's reasonable to do so. Special thanks also to Matthew Kinakin, Ewa Woźniak and Pinelopi Stylianopoulou for being my primary philosophical interlocutors and antagonists. And thank you to Thomas Lough whose patience this thesis tried the most.

I owe tremendous thanks to the Paisleys for providing me with unconditional love, encouragement, and support. Special gratitude goes to Ben for also being a daftie with me when I needed it.

My work has been significantly improved by the constructive criticism that I've received from a vast number of conference delegates and from anonymous referees at Philosophical Studies. I also gratefully acknowledge the financial support that I received for this research from the Scottish Graduate School for Arts & Humanities (2018 – 2022).

Finally, thanks to my examiners, Sara Bernstein and Stephan Leuenberger, for taking the time to engage with the thesis and provide useful feedback in the viva. Sara's papers initially got me thinking about causation and responsibility.

Author's Declaration

I confirm that this thesis is my own work and that I have: (i) read and understood the University of Glasgow Statement on Plagiarism, (ii) clearly referenced, in both text and the bibliography or references, all sources used in the work; (iii) fully referenced (including page numbers) and used inverted commas for all text quoted from books, journals, web, etc.; (iv) provided the sources for all tables, figures, data, etc. that are not my own work; (v) not made use of the works of any other student(s) past or present without acknowledgement. This includes any of my own works, that has been previously, or concurrently, submitted for assessment, either at this or any other educational institution; (vi) not sought or used the services of any professional agencies to produce this work; (vii) in addition, I understand that any false claim in respect of this work will result in disciplinary action in accordance with University regulations.

I declare I am aware of and understand the University's policy on plagiarism and I certify that this thesis is my own work, except where indicated by referencing, and that I followed the good academic practices noted above.

Published Material

At the time of the submission of this thesis (April 2023), portions of Chapters 2, 3 and 6, have been accepted for publication or published in the following journals:

Fearnley, L. (2022). 'Moral Worth, Right Reasons and Counterfactual Motives', *Philosophical Studies*

Fearnley, L. (forthcoming). 'Norms and Causation in Artificial Morality', *Proceedings to ACM conference in Intelligent User Interfaces*, (Association for Computing Machinery).

CHAPTER 1

Introduction

> We think of the world around us not as a mere assemblage of unrelated objects, events, and facts, but as constituting a system, something that shows structure, and whose constituents are connected with one another in significant ways. This view of the world seems fundamental to our scheme of things; it is reflected in the commonplace assumption that things that happen in one place can make a difference to things that happen in another in a way that enables us to make sense of one thing in terms of another, infer information about one thing from information about another, or affect one thing by affecting another. (Jaegwon Kim 1984, p. 153)

Central to Kim's idea of the interconnectedness of things is the notion of dependence. Things are connected with one another in whether something exists, or what properties it has, is dependent on, or determined by, what other things exist and what kinds of things they are. This thesis is about one such relation of dependency: that between causation and morality. Specifically, it is about the *inter*dependency between causation and morality. I aim to show that causal facts are partly determined by moral facts and that moral facts are partly determined by causal facts. By clarifying and evaluating this interdependency, I take myself to be illuminating our 'fundamental scheme of things' since it is in virtue of these dependency relationships that the world can be made intelligible, and by exploiting them we are able to manipulate the world in ways that express our agency.

It will be the purpose of this Introduction to provide an entry point into the concepts that will take centre stage in subsequent Chapters (causation and morality), and to set the stage for the critical work to come. Section 1 provides a state-of-the-art on how causation and morality are understood to relate to one and other in contemporary philosophical discourse. Sections 2 and 3 introduce the concepts of causation and morality, and demarcate the scope of the discussion for these concepts by outlining the kinds of moral and causal facts this thesis will be concerned with. Section 4 will provide an overview of the thesis and its conclusion. Finally, Section 5 will outline the aims of each individual chapter.

1.1 The Connection between Causation and Morality

The last couple of decades have seen an interest in the relationship between causation and morality, especially with regards to assessments of moral responsibility. Particular attention has been dedicated to the question of whether moral responsibility is grounded in or explained by actual causal sequences (Driver 2008), (Sartorio 2016), as well as whether an agent's causal contribution to a morally charged outcome can admit of degrees (Kaiserman 2018), (Demirtas 2022). Furthermore, there have been inquiries into the causal status of omissions and how omissions are supposed to fit into the landscape of responsibility (Sartorio 2004), (Whittle 2018). Causation's role in responsibility has also been explored with regards to assigning responsibility for harms brought about by a collection of agents (Kutz 2000), (Petersson 2013), in addition to what causal link, if any, is required to make one complicit in a harmful outcome (Kutz 2007). More generally, philosophers have examined the ways in which various complex features of causation bear on the understanding of moral responsibility (Bernstein 2016, 2017b).

Alongside normative ethics and metaethics, causation matters to the field of applied ethics. There has recently been a flurry of literature on the relationship between causation and individual liability for climate change (Sinnott-Armstrong 2005), (Garvey 2011), as well as causation's role in the ethics of war and self-defence (Beebee and Kaiserman 2018), (Sartorio 2019). Beyond moral philosophy, causation is essential in legal inquiry, as to establish legal liability in criminal and tort law one needs to show that the harm was brought about, induced or permitted by the agent (Hart and Honoré 1985), (Moore 2009), (Schaffer 2010). There is also emerging research in the field of experimental philosophy, where philosophers and psychologists have collaborated to analyse how our understanding of causal concepts shapes our moral judgements (Murray and Lombrozo 2017), (Grinfeld et al., 2020). Finally, causation matters to political theory, where the concepts of agency and responsibility are extended to collective entities (Pettit 1993, 2001), (Miller 2007). Examining whether states, institutions or international organisations could be responsible agents is an exercise which allows us to clarify the complexities of the political and policy-making process. If political values are distributed and shaped by the participants of the political process, and if these participants form part of the powerful collective entities, then the participants could be responsible for what comes of their significant causal power in the political domain.

It is fair to say that there is general acceptance that causation has a determining effect on morality. And, as the previous discussion indicates, there are a small contingent of metaphysicians who have sort to examine the nature of this relationship; unsurprisingly, most

have done so by trying to iron out the knots of causation's influence on the connection. However, there have been surprisingly few sustained expositions by moral philosophers on what exactly this determining relationship entails (notable exceptions include Driver 2008 and Clarke 2014). In particular, there is a lack of exploration into what kinds of causal facts make a difference to the truth value of moral assessments. This is especially surprising given the consensus surrounding causation's importance to morality. I suspect this lacuna has arisen because it is tempting for those working in ethics to remain neutral on questions regarding causation, lest she have to face up to dealing with the thorny issues such questions invite. However, sometimes it is only by getting involved in such issues that progress can be made. With this in mind, I aim to advance the debate by clarifying and evaluating what kind of causal facts determine different sorts of moral facts. I defend novel accounts of various moral assessments which include the explication of the causal conditions an agent must meet in order to be an appropriate target for such moral assessments.

We have seen how philosophers conceive of causation's effect in the moral domain. What about the other direction of travel? What does the contemporary literature say about morality's effect on causation? For a long time, philosophers working on causation did not consider that answers to the question 'when does one thing cause another?' could involve some kind of reference to morality. In the late 20th century dominate thinkers in causation such as David Lewis, J. L. Mackie and Wesley Salmon conceived of causation as a perfectly precise, fundamental, mind-independent metaphysical relation. To be picturesque about it, causation was (and by many still is) thought to comprise of a mind-boggling nexus of relations that binds its relata to form a fundamental structure of the world. To borrow Mackie's phrase, causation is the 'cement of the universe'. Given this perspective, it is no surprise that something so ostensibly nebulous and imprecise as morality was presumed not to be one of causation's determiners.

In the last couple of decades, however, philosophers, computer scientists, and statisticians have begun to question this assumption (Menzies 2004, 2007), (McGrath 2005), (Halpern 2008, 2016). The motivation for push back appears to come from a frustration at the lack of success in answering the question 'what is it for one thing to cause another?'. For the literature in the philosophy of causation has produced a wide variety of causal theories, but they in turn have produced a plethora of problematic cases that appear to refute them. This lack of success has prompted some to investigate other philosophical avenues to determining causal facts. One promising avenue has included an appeal to moral considerations, along with an appeal to a

variety of other normative features, the idea being that normative features play a decisive role in determining the truth value of causal facts.

These two perspectives have come to dominate the causation literature. Yet, there is a lack of clarity regarding exactly what these two approaches to causation entail. In this thesis, I aim to advance the philosophical research into causation by delineating the central components and metaphysical commitments of each of these perspectives, in addition to framing the boundaries of discussion between them.

1.2 Scope Setting: Causation

Having broadly outlined how causation and morality are understood to relate to one another in contemporary philosophical discourse, I will move to introduce the two concepts individually and in more detail. I will also set the scope of the upcoming analysis by outlining what kind of causal and moral facts this thesis will be concerned with. I begin with causation.

Questions about cause and effect are pervasive in our everyday lives. These range from the quotidian to those of huge significance. Who caused the house plant to die? Why isn't the computer turning on? Did your lateness cause your friend's frustration? Did the doctor's neglect cause the patient's death? Given the centrality of questions like these, it is no surprise that there have been many attempts to theorise about causation both within and outside of philosophy. Philosophical concern with this topic dates back to at least Plato and Aristotle and continues to occupy a central role in the doctrines of philosophers in the early modern period, including Descartes, Locke, Hume and Kant. Outside of philosophy one finds a rich and extensive literature in statistics, law, cognitive psychology, computer science and biology on how best to understand the notion of cause and effect as well as causal inference.

In addition to the vast array of disciplines in which causation has been studied, there are also a wide range of research questions one can engage in when theorising about causation. For example, one could answer research questions concerning causal cognition; roughly, the study of how individuals gain causal knowledge through detecting, learning, and reasoning from statistical regularities. Or one could seek to investigate how people actually talk about causation and causal concepts. Such an investigation typically consists in analysing the empirical evidence that results from experiments involving human participants and their reaction to a range of hypothetical causal cases. In this thesis, I do not seek to engage in any of these research topics. My starting point for theorising about causation will follow the central

question that has occupied metaphysicians working on the subject: what is it for one thing to cause another? This is a question about causal facts of the kind '*c* caused *e*'. Specifically, it is a question about what kind of relation makes claims like '*c* caused *e*' true, and in virtue of what does this relation obtain. Examining these questions and their answers will be the focus of Chapters 2 and 3.

The metaphysical literature has produced plenty of diverse proposals to determine facts of the kind '*c* caused *e*'. Causal relations have been reduced to regularities (Mackie 1974), physical connections (Dowe 1992, Salmon 1994), probabilities (Ellery 1991), dispositions (Molnar 2003), mechanisms (Glennan 1996, 2009), and agency (Menzies and Price 1993). But the clear favourite is an approach that sees counterfactual dependence as the key to such relations. It is a safe bet that David Lewis's ground-breaking paper 'Causation' (1973) remains the most well-known iteration of the counterfactual approach to causation. According to this treatment, '*c* caused *e*' is true if it is the case that *c* and *e* are distinct events, both *c* and *e* occur, and had *c* not occurred *e* would not have occurred. Lewis's analysis of causation has had a tremendous impact on the philosophical literature and its success no doubt stems from its relative simplicity and its ability to match common sense in a wide range of ordinary cases. In light of its success this thesis will limit itself to focusing on counterfactual treatments of causation. Specifically, it will focus upon Lewis's 1973 counterfactual analysis, in addition to a more recent explication of the counterfactual approach presented by James Woodward in his 2003 book *Making Things Happen: A Theory of Causal Explanation*. Woodward's theory, which some term 'interventionism', roughly says that *c* is a cause of *e* where there exists an intervention or manipulation of *c* that would bring about an associated change in *e*. *Making Things Happen* has already made a substantial contribution to the subject of causation and causal explanation, a contribute that will no doubt be a lasting one. Indeed, I think it's fair to say that Woodward's interventionism has now supplanted Lewis's original analysis in terms of being the authoritative counterfactual based causal theory. I suspect this is because interventionism is formed of an extremely rich and complex theoretical framework which allows the account to overcome the faults faced by Lewis's elegant yet simple view. Chapter 3 will be dedicated to an analysis of how interventionism determines facts of the kind '*c* caused *e*'.

In addition to theorising about facts of the kind '*c* caused *e*', it is becoming increasingly popular for philosophers to categorise and analyse the different properties that a causal relation can possess. On the face of it, investigations into what properties a causal relation can possess seem to be independent of what conditions are required to make '*c* caused *e*' true. That is, we can

debate what *else* can be true of causal relations without first settling the question of what it takes for something to be a cause. As the thesis progresses, I will move to explore some of these properties. In particular, I will investigate the property of "causal stability" and "causal robustness". What I call causal stability was first introduced by David Lewis (1987) under the name "sensitivity" (p. 184). Since Lewis the concept has been left relatively neglected until recently where philosophers such as Woodward (2006) and Laura Franklin-Hall (2016), as well as psychologists and philosophers such as Nadya Vasilyeva, Thomas Blanchard and Tania Lombrozoa (2018) have begun to properly analyse the concept. In this thesis, I define causal stability roughly as the extent to which a causal relationship would continue to hold across different background conditions when we hold the occurrence of the cause fixed.

The authors who deploy the concept of causal stability often do so by running it together with the concept of causal robustness. Stability and robustness are taken to refer to the same phenomena in the literature, and are thus deployed as interchangeable terms. However, in this thesis I do not use these terms interchangeably. I assign different meanings to these concepts, and I propose an account of how to capture each respectively. I suggest that these are two distinct properties that a causal relation can possess, and that they therefore offer up different kinds of causal information. I define a causal relationship as robust to the extent that the causal relationship would continue to hold across different background conditions when we do *not* hold the occurrence of the cause fixed. I provide a more detailed explication of the difference of these two concepts in Chapter 6.

So, to clarify the scope in terms of causation: this thesis will begin by focusing on facts of the kind '*c* caused *e*', and will later also focus on facts about causal stability and causal robustness.

1.3 Scope Setting: Morality

Whilst the metaphysics of causation is concerned with how the world *is*, morality is largely concerned with how we are to act in it. There is a kaleidoscopic array of entry points from which one can begin to engage in moral theorising, so many that it is difficult to summarise just what theorising about morality entails. Let me narrow the field somewhat. The second half of this thesis will focus on two distinct kinds of moral theorising; namely, theorising about facts pertaining to moral responsibility and moral praiseworthiness.

Chapter 5 will focus on moral responsibility. Making judgements about whether a person is morally responsible for her behaviour, and holding others and ourselves responsible for that behaviour, forms a fundamental and familiar part of our moral practices and our interpersonal

relationships. Judging that a person is morally responsible for her conduct involves attributing a host of agency-related capacities to that person, such as motivation, control, foreseeability, freedom and epistemic competence. It also involves viewing that behaviour as arising from an exercise of these capacities. There's an abundance of literature canvassing what exactly these capacities amount to. Prominent approaches include the reactive attitude view first defended by F. Strawson (1962), the reason-responsiveness view defended by the likes of John. M Fischer and Mark Ravizza (1998) as well as David Brink and Dana Nelkin (2013), and the quality of will view endorsed by folks such as Susan Wolf (1990) and Nomy Arpaly (2003). I won't be interested in adjudicating between which of these approaches best captures the capacities required for moral responsibility — in fact, I largely set aside the question of what capacities makes one morally responsible. Instead, I will concern myself with what kind of powers, specifically causal powers, are required for moral responsibility.

Chapter 6 will focus on moral praiseworthiness, or as it is referred to in the Kantian literature "moral worth". Moral worth can be defined as a particular way in which we find an action valuable. It is thought that this value is largely derived from the agent's motive for acting such that an action has moral worth to the extent that it is issued from an appropriate motive. The chief aim for theories of moral worth is to explicate what these appropriate motives consist in. Prominent theories take their departure from the Kantian idea that morally worthy actions are issued from the motive of duty (Herman 1981), (Stratton-Lake 2000), whilst others emphasise the importance of being motivated by the knowledge that one is doing the right thing (Sliwa 2016). I examine a third popular view of moral worth defended by Nomy Arpaly (2002) and Julia Markovits (2010). According to this account, a right action is praiseworthy if and only if the agent performed the action in response to the right reasons, that is, the reasons which make performing the action the right thing to do. The view sets itself up as a rival to traditional Kantian and moral knowledge theories, and has therefore attracted many contemporary sponsors who find these ways of understanding moral worth unpersuasive. Chapter 6 analyses and improves upon Arpaly and Markovits's account of moral worth.

In addition to these two specific categories of moral facts, in Chapters 2 and 3 of the thesis I call upon a plethora of normative considerations in my discussion of causation. Some of these normative considerations have moral import; they pertain to facts about how people ought to conduct themselves in accordance with morality. Examples of distinctly moral normative considerations are expressed by statements like 'you should keep your promises', 'you ought to treat others with respect', and 'you shouldn't lie'. Whilst I draw on these normative

considerations in my discussion, it is worth noting that I do not provide an account of how to determine them as I do with facts about moral responsibility and moral praiseworthiness. Rather, I merely take for granted that the moral claims expressed by statements such as these are true.

So, to clarify the scope in terms of morality: this thesis will begin by focusing on the heterogeneous category of normative considerations including those normative considerations with distinctly moral import. Then Chapters 5 and 6 concern themselves with facts about moral responsibility and facts about moral praiseworthiness.

1.4 Overview of Aims

Now that I have identified the kinds of moral and causal facts I'll be interested in, let me outline in broad terms how these facts are supposed to relate to one and other.

To reiterate, the aim of this thesis is to argue for the interdependency between causation and morality. In the first part of the thesis, I explore the question: are causal facts determined by moral facts? In the second part of the thesis, I explore the question: are moral facts determined by causal facts?

In regards to the first question, I argue that moral facts partly determine causal facts. In particular, I argue that an appeal to facts about morality, and normative notions more generally, can explain a vast range of seemingly disparate and unusual features of the causal concept. Specifically, I argue that the addition of a normative framework equips a causal account with the resources to meet various desiderata associated with theories of causation. Insofar as a causal theory is able to satisfy these desiderata, I suggest that causal accounts that incorporate normative frameworks are more successful than ones which do not.

In regards to the second question, I argue that causal facts partly determine moral facts. In particular, I argue that certain causal facts determine facts about whether one is morally responsible for an outcome, and whether and to what extent one is praiseworthy for that outcome. With regards to assessments of moral responsibility, I argue that whether one can be held morally responsible for the effects of an omission depends upon whether one's omission enjoys a stable causal connection with the outcome for which we want to attribute moral responsibility. And with regards to assessments of praiseworthiness, I argue that whether and to what extent one is praiseworthy for doing the right thing depends upon how causally robust one's motivation was for acting.

1.5 Chapter Summaries

With the argument spelled out in broad terms, let me give an overview of how I will make this argument by running through the structure of the forthcoming discussion in more detail. In Chapter 2, I introduce two 'meta-causal' approaches to causation that have come to dominate the philosophical discussion. The first approach, which I term "causal realism", says that causation is a mind-independent structural feature of the world. According to causal realism, causal facts of the kind '*c* caused *e*' exist, and they do so regardless of what anyone happens or say or think about the matter. The second approach, which I term "normativism" argues that moral facts, and normative notions more widely, play a central role in determining causal facts of the kind '*c* caused *e*'. Contemporary philosophical discourse presents these two doctrines as incompatible approaches to causation. One of the chief ambitions of Chapter 2 is to clarify and evaluate whether the prominent discourse is correct in conceiving of causal realism and normativism as incompatible.

Ultimately, I argue that they are to some extent compatible; the causal realist can draw upon a limited set of normative considerations to determine what causal relations there are in the world without undermining her metaphysical commitments. My argument in this Chapter has significant implications. For one thing it opens the door to developing a novel 'hybrid' view of causation which preserves a realist metaphysics whilst appealing to a restricted set of normative considerations. The second ambition of this Chapter is to evaluate how this new hybrid view fares when compared to a normativist view that is not restricted in the set of normative considerations it can draw upon. I conclude by arguing that an account which is not restricted in its incorporation of normative considerations can better satisfy several desiderata associated with theories of causation, and therefore provides a more successful approach.

Having laid the groundwork in Chapter 2, Chapter 3 goes onto to analyse whether one particular account of causation is compatible with causal realism. The account I focus on is James Woodard's (2003) interventionism. There is a small but growing debate about whether interventionism is compatible with causal realism. The debate centres around interventionism's incorporation of causal models which it deploys to determine causal facts of the kind '*c* caused *e*'. Some philosophers, believe that in order to construct correct causal models, interventionism needs to appeal to moral facts and normative considerations more broadly (Hitchcock 2007) (Halpern 2016). Others, like Woodward himself, believe that models can be correct in the absence of such considerations. My aim in this Chapter is to provide reasons in favour of

thinking that a successful interventionist theory is one that relies on moral facts and other normative notions when constructing causal models. Specifically, I argue that interventionism must call upon the kinds of normative considerations which are incompatible with a realist conception of causation.

Together Chapters 2 and 3 examine morality's effect on causation. Chapters 5 and 6 move to look at causation's effect on morality. To provide a smooth transition between these two lines of inquiry, Chapter 4 readies the reader for what's to come by mapping the way forward and briefly clarifying the dietetic.

Chapter 5 begins with two plausible claims about moral responsibility. First, an agent is morally responsible for a particular outcome only when she is causally responsible for that outcome. Second, agents can be morally responsible for outcomes that result from omissions. I then point out that those who endorse both of these claims face a pivotal challenge; namely, to accommodate for the idea that agents can be causally responsible for outcomes through their omissions. That is, they must provide an account of omissive causal responsibility. The primary aim of Chapter 5 is to provide such an account.

My account of omissive causal responsibility appeals to the concept of causal stability. I will suggest that an agent is omissive causally responsible for an outcome when the causal connection between her omission and the outcome is a relatively stable one. Whereas an agent is not omissive causally responsible for an outcome when the causal connection between her omission and the outcome is a relatively unstable one. In this way, I demonstrate that causal facts determine facts about moral responsibility. Along the way, I will examine and reject two other candidate theories for omissive causal responsibility. I will also advance the debate by fixing the target and setting the success conditions for theories of omissive causal responsibility.

Next, in Chapter 6, I will explore another way in which causal facts determine moral facts. This time the types of moral facts I'm interested in are assessments about moral praiseworthiness, and the types of causal facts I'm interested in concern facts about causal robustness. I will argue that whether and to what extent one is praiseworthy for doing the right thing depends upon the causal robustness of one's motivation. I argue for this in the context of responding to Arpaly (2002) and Markovits's (2010) theory of moral worth, according to which a right action is worthy of praise if and only if the agent performed it in response to the right-making reasons. I first argue that this view not as successful as contemporary discussions suggest because it

fails to adequately satisfy two important desiderata associated with theories of moral worth. Next, I argue that the view can meet these desiderata once the theory attends to facts about the causal robustness of the agent's motive. Hence, this Chapter argues that facts about causal robustness determine facts about moral praiseworthiness.

Finally, in Chapter 7, I conclude the thesis by examining one potentially significant implication of what has been argued thus far. The overarching aim of the thesis is to demonstrate that moral facts partly depend upon causal facts and that causal facts party depends upon moral facts. If this is true, then we might be met with a potential circularity of dependency between causality and morality. In the conclusion, I demonstrate why my argument does not entail such a circularity.

CHAPTER 2

Causal Realism and Normativism

2.1 Introduction

Within the last two decades one view about causation has begun to gain serious traction. According to this view, which we might call "normativism", normative considerations play a central role in determining causal facts. By normative considerations I mean claims or facts involving normative concepts, in particular claims or facts about normality and abnormality. The basic idea is that the normative status of an event makes a difference as to whether that event enters into causal relations or not. For instance, events which are considered to be normal are said to be causally inert, whereas events which are considered to be abnormal are said to be causally efficacious. Proponents of the normative approach argue that incorporating normative considerations into a causal theory has immediate intuitive appeal, but the view's strongest justification is its success in explaining a vast range of seemingly disparate and unusual features of the causal concept. In particular, it seems like the addition of a normative framework equips a causal account with the resources to meet various desiderata associated with theories of causation. A combination of the position's immediate intuitive appeal and its various theoretical virtues has seen normative approaches enjoy a surge in popularity.

Normativism isn't the only game in town, however. According to a long-standing, widely accepted thesis we might call "causal realism", causation is a mind-independent structural feature of reality. Causation is comprised of a mind-boggling nexus of relations that exists as a feature of the external world. According to this doctrine, causal facts exist, and they do so regardless of what anyone happens or say or think about the matter.

We might think of causal realism and normativism as 'meta-causal' views. Unlike 'on the ground' theories of causation, like the counterfactual analysis and regularity accounts, realism and normativism are not in the game of specifying the necessary and sufficient conditions for causal concepts per se. Rather, normativism and realism are the umbrellas under which such

theories sit. They propose high-level theorising about the wider systemic features of causation, the delineations of which can bottom-out in various and diverse ways.

Contemporary philosophical discourse conceives of these meta-causal views as incompatible approaches to causation. Indeed, it's common for philosophical discourse to frame normativism and realism as *opposing* causal doctrines. To put it simply, the idea is that what is normative is determined by minds and perspective, hence normative considerations are understood to be mind-*dependent*. The causal realist, therefore, is not entitled to deploy normative notions to determine what causal connections there are in the world, lest those connections themselves be rendered mind-dependent. Let us refer to the claim that causal realism is incompatible with an appeal to normative considerations as INCOMPATIBILITY.

INCOMPATIBILITY: causal realism is incompatible with a normative approach to causation.

INCOMPATIBILITY has high stakes implications for the causal realist. The realist would be missing out on something tremendously philosophically valuable if they cannot incorporate norms to determine the causal facts, since supplementing one's causal theory with a normative framework is said to yield huge theoretical benefits. For this reason, I think INCOMPATIBILITY demands a lot more scrutiny than it has thus far received. In particular, it's worth paying more attention to the arguments given in favour of INCOMPATIBILITY, as well as exploring the ways a realist might resist these lines of argument. This is precisely what I will do in this Chapter. I aim to answer the question: is causal realism compatible with a normative approach to causation? Specifically, I will examine whether the causal realist can supplement her preferred account of causation with a normative framework without undermining her metaphysical commitments regarding what she takes causation to be — that is, without undermining the idea that causation is a mind-independent feature of the world.

Ultimately, I'll argue that the realist can invoke a restricted set of normative considerations into her causal theory. To argue for this, I will show that not all normative considerations are mind-dependent, some are mind-independent, the realist is justified in incorporating these mind-independent normative considerations when determining causal relations without threating her realist credentials. This Chapter therefore pushes back against the prevailing philosophical discourse by arguing that INCOMPATIBILITY is false.

This conclusion has significant upshots. For one thing it opens the door to exploring a new 'hybrid' view of causation, one which secures a realist conception of causation whilst determining causal relations through a normative framework. The final Section of this Chapter assesses how such a hybrid view fares when compared to a non-hybrid normative view. That is, a normative view which does not aim to secure the realist conception of causation, and is therefore unconstrained in the types of normative considerations it can draw upon. I argue that a non-hybrid view is more successful at satisfying the seemingly disparate desiderata associated with theories of causation than a hybrid view. Insofar as this is the case, I argue that it provides a more successful approach to causation, even if it comes at the expense of conceiving of causation as a mind-independent relation.

Roadmap: In Section 2, I sketch out the core elements of causal realism. In Section 3, I do the same for the normative approach. In Section 4, I outline three desiderata associated with theories of causation, as well as demonstrating the ways in which normativism satisfies such desiderata. In Section 5 I clarify and analyse the arguments in support of INCOMPATIBILITY. Next, in Section 6, I argue against INCOMPATIBILITY. Then in Section 7, I provide reasons for thinking that a non-hybrid form of normativism will deliver a more successful theory of causation than a hybrid form of normativism. Finally, I offer some concluding remarks in Section 8.

2.2 Causal Realism

There is no hard and fast definition of causal realism, but there are two metaphysical assumptions that seem to be central to all iterations of the realist position. First is a claim about existence. Causal relations are said to exist, and they do so as a structural feature of the world. To be picturesque about it, the structure consists of an incomprehensibly huge, homogeneous web of relations which, when seen together from a God's eye view, represent the entire causal history of the universe. According to Peter Menzies (2009), realists suppose that causal relations are of a distinct kind; namely, *natural* relations whose relata are *events* (p. 342).[1] Broadly speaking, natural relations are external relations (relations that are not fixed by the features of their relata), such as spatial and temporal distance, that play a central role in the scientific conception of reality. Attempts to understand natural relations look to carve nature at

[1] Menzies does not explicitly ascribe this view to the "causal realist". Nonetheless, it's clear that he has something akin to causal realism in mind.

its joints in the sense that the investigation picks apart the skeletal structure of reality with a view to better understanding those kinds made in or found in nature.

The second metaphysical assumption of realism concerns independence. The fact that a certain causal relation exists is independent of anyone's intentional, epistemic or doxastic states (except, of course, when the causal connections concern such states). Every one of these causal relations obtains independently of our ability to discover that they do. Mind-independence is frequently taken as a criterion or necessary condition for realism about a phenomenon. To say that something is mind-independent is thought to be at least part of what it is to say that it is real, objective, truly exists, and that it is not just a product of our imaginations like unicorns or dementors.

Jaegwon Kim articulates causal realism as follows:

> [A]ccording to causal realism every event has a unique and determinate causal history whose character is entirely independent of our representation of it. We may come to know bits and pieces of an event's causal history, but whether we do, or to what extent we do, and what conceptual apparatus is used to depict it, do not in any way affect the causal relations in which events stand to other events. (1988, p. 230)

The idea that whatever the causal relation turns out to be it exists independently of how we choose to conceptualise of it, is strongly defended by David Lewis. In the below passage, Lewis draws a sharp distinction between the concept of causation and what we're inclined to *say* about the concept of causation:

> We sometimes single out one among all the causes of some event and call it "the" cause, as if there were no others. Or we single out a few as the "causes," calling the rest mere "causal factors" or "causal conditions." Or we speak of the "decisive" or "real" or "principal" cause. We may select the abnormal or extraordinary causes, or those under human control, or those we deem good or bad, or just those we want to talk about. I have nothing to say about these principles of invidious discrimination. I am concerned with the prior question of what it is to be one of the causes (unselectively speaking). My analysis is meant to capture a broad and nondiscriminatory concept of causation. (1973, p. 559)

Lewis's ambition permeates many contemporary causal accounts insomuch as the majority of realists are happy to concede that our causal *talk* is influenced by 'principles of invidious discrimination', but the "true" nature of causation is considered unaffected by such capricious judgements. Indeed, the enterprise of understanding how and why humans distinguish between causal concepts is often viewed as a separate project, one which only serves to obscure the metaphysical underpinnings of the causal relation.

To illustrate the realist position in action, consider a token instance of the causal relation expressed by the following claim: 'the lighting strike caused the forest fire'. Call this instance of the causal relation *lightening*. For a causal realist, the world contains countless instances of causal relations like *lightening*. Further, the truth conditions for causal claims, such as 'the lighting strike caused the forest fire' just are the conditions for *lightening*.[2] That is, what makes a causal claim true is just the existence of the instance of the causal relation; truth value does not depend upon anything outside of the causal relation itself.

The majority of philosophers writing about causation claim to be causal realists.[3] Traditionally, the chief challenge for the realist has been to identify what exactly the causal relation is comprised of. Popular strategies have argued that the causal relation is grounded in or supervenes on particular structures and patterns, though they differ with respect to what these structures and patterns are. For example, regularity theorists like J. L. Mackie (1980) argue that causal relations are constituted by patterns of regularities. Probabilistic theories like the one defended by Ellery Eells (1991) state that causal relations are relations of probabilistic dependence between events. Whilst process theories, like those advanced by Phil Dowe (2000), state that causal relations reduce to causal process that involve conservation or transference of physical quantities. The most influential analysis, championed by Lewis (1973), takes causal relations to be relations of counterfactual dependence. Though these accounts provide diverse answers, they all aim settle the same ontological question — what *is* the causal relation?

[2] This is how Sara Bernstein (2017a) illustrates causal realism (p. 223).
[3] I think it's more difficult than one might first suppose to find philosophers who unambiguously advocate for a realist conception of causality. Still, Menzies (2009, p. 342), Bernstein (2017a, p217), and Halpern (2016, p.108) note that most philosophers take themselves to be giving an analysis of an objective, mind-independent causal relation.

Though causal realism continues to be philosophical orthodoxy, it has not gone unchallenged. Most famous is David Hume's critique of the idea that there must be a "necessary connection" between causes and their effects. For Hume the only type of reasonable account of causation is one in which the power of causes rests wholly in the human mind; not in anything external to the mind. In *A Treatise of Human Nature,* Hume states:

> The necessity or power, which unites causes and effects, lies in the determination of the mind to pass from one to the other. The efficacy or energy of causes is neither plac'd in the causes themselves, nor in the deity, nor in the concurrence of these two principles; but belongs entirely to the soul, which considers the union of two or more objects in all past instances. 'Tis here that the real power of causes is plac'd, along with their connexion and necessity. (1.3.14.23)

Although there is debate as to what Hume really meant in this passage and others like it in the *Treatise*, according to most contemporary taxonomies of causation, Hume was a causal irrealist par excellence. He took causation to be a product of human mental activity rather than a natural relation to be found in an external reality. Hume had some illustrious followers. Bertrand Russell (1912) for example, ridiculed causation as "a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm" (p. 1). Unlike Hume, Russell did not attack realism on the grounds that causation could exist outside of the mind. Rather, Russell attacked the idea that causation is amongst the most basic constituents of a scientific reality. Russell argued that there is no metaphysical account of causation compatible with the completeness of advanced physics and recommended the "excision" of the word "cause" from philosophical vocabulary, seeing it as inextricably bound up with misleading connotations. I do not come down on either side of the realist/irrealist debate here; I mention Hume and Russell only to further clarify what the realist position is (or rather what it is not).

2.3 Normativism

Hart and Honoré (1985) are often credited as originators of the normative approach.[4] When discussing the role causation plays in the law, they contend that a cause should be understood as an intervention, analogous to a human action, that makes a difference to the way things

[4] For example, Menzies (2009, p. 355) credits Hart and Honoré as one of the first to make poplar the idea that causation is infused with norms.

would *normally* develop. According to Hart and Honoré, common experience teaches us that left to themselves things have dispositions or characteristic ways of behaving. These things thus persist in states different from those which we bring about through our manipulations. A cause, they argue, is something which interferes with the characteristic way things normally behave. They put the idea as follows:

> When we assert that A's blow made B's nose bleed or A's exposure of the wax to the flame caused it to melt, the general knowledge used here is knowledge of the familiar way to produce, by manipulating things, certain types of change which do not normally occur without our intervention. If formulated they are broadly framed generalizations, more like recipes, in which we assert that doing one thing will 'under normal conditions' produce another. (1985, p. 31)

By defining causation in terms of interventions on what would normally occur, Hart and Honoré thereby introduce a novel idea into the causation literature; namely, that causal relations are sensitive to what are often referred to as normative considerations.

I'll return the notion of normality shortly, for now we can characterise the broader category of normative considerations as roughly involving claims or facts about normativity. Claims about normativity can express evaluations or judgements about values, asserting that something is good, bad, desirable, justifiable, impermissible, would be an example of a normative statement. Perhaps somewhat confusingly, claims about normativity can also relate to descriptive standards, these claims do not involve evaluations but rather reports or observations about the way things actually are. Contemporary normative accounts of causation appeal to a wide variety of evaluative and descriptive normative considerations, but given how hugely heterogenous these categories are, it would not be feasible to offer an in-depth analysis into the ways in which all kinds of normative considerations are supposed to determine causal relations (at least not here). Consequently, I'm going to look at one particular, though multifaceted type of normative consideration which has dominated the literature since Hart and Honoré — normality.

2.3.1 Normality

Many philosophers, including Menzies (2004, 2007), Hitchcock (2007), Hall (2007) and Halpern (2016) have begun to invoke normality into their theories of causation. Despite its widespread use however, there has been no sustained analysis of what is entailed by the concept (at least not to my knowledge). Though this is not especially surprising. Normality is ambiguous; it is a concept that is comprised of many other distinct and often competing normative considerations, making it difficult to provide a clean analysis. Indeed, discussion of the concept outside of causation has also been fairly limited.[5] With all that said, it appears to me that behind the diversity of normativist views of causation, there hides a shared and consistent understanding of normality, albeit one that is rarely explicitly articulated.

Normality is typically understood to be constituted by a plethora of more specific norms. These norms are both prescriptive and statistical. To say something is a statistical norm is to say that it conforms to a statistical mode. For example, it is statistically normal for Glasgow to have more rainfall than Milan during the winter, so if Glasgow were to have less rainfall than Milan one winter, Glasgow's weather would violate a statistical norm. By contrast, to say something is a norm in a prescriptive sense is to say that that thing follows a prescriptive rule. These rules are constituted by the way things *ought* to be or are *supposed* to be. Prescriptive norms can take many forms. Some norms are moral; for example, it's generally believed that people are supposed to keep their promises, even if there are no explicit laws or rules demanding this behaviour. There are also norms of etiquette which establish standards of how people are supposed to act in certain social contexts. Laws too can create norms which establish expectations about how people regulate their behaviour in societies. Policies enacted by institutions can be norms; for example, a company may have a dress code creating expectations around how employees dress for work. As McGrath (2005) points out, there are also norms of proper functioning for organisms and machines. Alarm clocks are supposed to ring at their set times and human hearts are supposed to pump blood around the body, and there's a sense in which 'supposed to' has normative force here - failure to function properly is a failure to meet a certain kind of standard.

[5] A notable exception is the philosophy of biology where there have been serious attempts at analysing the concept of normality. For example, Chadwick (2016) examines the concept of normality with a particular focus on its characterisation and uses in biology.
Elsewhere, Catita, Águas and Morgado (2020) analyse and categorise competing notions normality used in clinical medicine.

Furthermore, there also seems to be an exceptional sense in which the way things ought to be, and interestingly this expectational sense is entailed by a statistical, rather than prescriptive norm. For instance, imagine that Professor X uses room 101 every Friday afternoon to teach her biology class. In doing this regularly, she creates a sense in which room 101 'ought' to be booked this Friday, and this 'ought' is not a prescriptive one but rather one that is entailed by our expectations of what is statistically likely to occur in the future.

One of the reasons why the concept of normality is ambiguous is because sometimes these specific kinds of norms can pull in different directions; something could be a norm according to one standard but not according to a different standard. For example, it could be the case that keeping promises is normal according to a moral standard, but it could nonetheless be true that people statistically do not keep their promises. When normativists deploy the concept of normality, I take them to be weighing up these kinds of considerations against each other to deliver an overall evaluation of what's all things considered most normal. So, one might say that keeping one's promises is normal because all things considered keeping promises is more normal than not keeping promises. Normativists seem to use the notion of normality as a heuristic or approximation for what we expect will occur based upon the standards set by a multiplicity of prescriptive and statistical norms.

Alongside normality, normativists also appeal to the notion of abnormality. A set of circumstances or event is abnormal to the extent that it violates a prescriptive or statistical norm. For instance, if Glasgow were to have less rainfall than Milan one winter, then all other things being equal, Glasgow's weather would be abnormal. Similarly, if an alarm clock were to fail to ring at its set time, then that failure would constitute a violation of a functional norm, and would thus be abnormal. Much like normality, judgements about whether and to what extent something is abnormal can be equivocal when norm violations pull in different directs. And as with normality, I take the normativist to be weighing up these competing considerations in order to deliver an approximate, overall judgement about the extent to which something fails to align with our expectations of what would normally occur. [6]

[6] At this point one might be concerned that the concept of normality remains too ambiguous to delivery clean, clear causal verdicts. The worry might go that there will be cases where it's ambiguous as to whether *c* is normal or not, and it will therefore be ambiguous as to whether *c* is a cause. I believe this worry has been inadequately addressed by defenders of normativism. Though some authors such as Hitchcock and Knobe (2009) and Sarah McGrath (2005) have said something specific

It is worth pointing out that in addition to normality and abnormality, sometimes philosophers working on causation also refer to notions of "defaults" and "deviations". I understand these sets of terms to be roughly equivalent: a default refers to a normal state of affairs, whilst a deviation refers to an abnormal state of affairs. Going forward I will deploy both kinds of terminology and treat them as if they were synonymous sets of concepts.[7]

2.3.2 A Normative Account Sketched

With the conceptual resources spelled out in a little more detail, we're now in a position to see how advocates of the normative approach incorporate ideas of normality and abnormality into their causal theories. The basic idea is that the claim '*c* caused *e'* depends, amongst other things, on whether *c* represents a default or normal occurrence. More specifically, *c* is a cause if *c* represents a deviation from what we would normally expect to occur, conversely *c* is not a cause if *c* represents a default or normal state of affairs.

I say that '*c* caused *e'* depends upon *c's* normative character *and 'other things'* because an event's being normal or abnormal is not sufficient to grant it causal status. The event in question could be an abnormal event but it also needs to be *connected* to another event for it to enter into a *causal* relation. That is, the abnormal event must bear the appropriate metaphysical relation to the effect event for it to be a cause of that effect. Hence, in addition to a condition about the event's normative character, a normative account will also require a condition that establishes whether the event has the appropriate metaphysical relation to the effect. There are many of these conditions offered in the literature, but the one most commonly used by

about what counts as 'normal' which might mitigate this worry. I say more about this in the main text in Section 5.

[7] There's a case to be made that contra to what I'm saying here, the concepts of defaults and deviations are not synonymous with the concepts of normality and abnormality. Support for the idea that these are distinct notions could come from the likes of Halpern (2016), who provides a definition of defaults which does not invoke normality. Halpern says: "A default is an assumption about what happens, or what is the case, when no additional information is given. For example, we might have as a default assumption that birds fly. If we are told that Tweety is a bird and given no further information about Tweety, then it is natural to infer that Tweety flies. Such inferences are defeasible: they can be overridden by further information" (p. 77). Given this characterisation, one might think that defaults do not equate to what's normal, and can therefore do extra work when it comes to determining causal claims. However, it seems to me that this characterisation of a default can be reduced to talking in terms of normality. Defaults are based upon what is statistically normal. Adopting the assumption that Tweety can fly is a successful inference only insofar as it's statistically normal for birds to fly. Halpern himself notes of this connection later in the passage (p. 78).

normativists is the counterfactual test or variations thereof. The counterfactual test for causation, championed by David Lewis (1973), is the most influential test for causation in metaphysics. According to the simplest version of the view, some event *c*, is a cause of another event, *e*, when there is counterfactual dependence between *e* and *c*, such that if *c* had not occurred, then *e* would not have occurred. To illustrate, suppose that Suzy throws a rock at a window and the window shatters. Suzy is a cause of the shattered window, according to the simple counterfactual test, because it is true that had Suzy not thrown the rock, the window would not have shattered. There is counterfactual dependence between the shattered window and Suzy's throwing of the rock. When we pair the counterfactual test with a condition regarding an event's normative character, we get something like the following: *c* is a cause of *e* if *c* represents a deviation from what we would normally expect to occur, and *e* counterfactually depends upon *c*. Had *c* represented a default or normal state of affairs it would not qualify as a cause of *e*.[8]

This is only a rough sketch of a normative view that I have provided for illustrative purposes, as far as I'm aware no one advocates for this exact view. This is no doubt because the simple counterfactual test described above has been surpassed by more sophisticated counterfactual tests which are supposed to side-step the faults of Lewis's original, and normativists typically invoke these more sophisticated variations. But the details need not detain us here, for the aim of this Chapter is not to analyse any one particular normative theory, rather the aim is to analyse the broader claim that unites such theories — that normative considerations determine causal facts. Still, the above sketch is certainly in the spirit of normative accounts, and it will do for the purposes of this Chapter.

Let's see the view in action then. Suppose that you promise to feed my fish when I'm on holiday, but you fail to do so, and the fish dies. The above sketch says that you are a cause of the fish's death because there is counterfactual dependence between the fish's death and your failure to act; had you fed the fish, the fish would not have died. And your failure to keep your promise is a deviation from a moral norm, and this explains why you and *not anybody else* is a cause of the fish's death.

[8] Much like Lewis's original counterfactual analysis of causation, the above sketch can be seen as a sufficient condition for causation. Promoted to a sufficient and necessary condition wouldn't do; it is easy enough to have circumstances in which *c* causes *e*, even though *c* is a normal thing to have happened. But as a sufficient condition on causation, it has struck many as exactly right, and therefore as a promising starting point for a full-blown analysis of causation.

Normativism is thought to have immediate intuitive appeal which we can see with the example just outlined. Although every person failed to feed my fish, intuitively the only person causally implicated in the fish's death is you, and this appears to be because you, and no one else, promised to do so. An account which appeals to moral norms and deviations thereof in order to identify causal relations is therefore able to make sense of the intuition that you are a cause of the fish's death. More broadly, the notion that a cause is something that deviates from what we expect to occur is for many a very plausible starting point for theorising about causation. It reflects Hart and Honoré's thought that we interact and manipulate the world around us by disrupting the way things normally behave. With that said, normativism's ultimate justification is said to come from the success it has in explaining a vast range of seemingly disparate and unusual features of the causal concept. In particular, incorporating notions of normality and abnormality into a causal analysis appears to satisfy many desiderata associated with theories of causation. These include: the ability to distinguish between background conditions and causes, the ability to identify which omissions participate in causal relations, and the ability to better vindicate our intuitive causal judgements. Once more, normative views seem to satisfy these desiderata in a more comprehensive and unified way than non-normative views.

To illustrate and support this argument, I'm going to spend a little time outlining the ways in which normative considerations are supposed to meet the three desiderata just described. Before I begin, I want to emphasise three things. Firstly, I want to emphasise that the desiderata I consider do not represent an exhaustive list. I will consider three pieces of evidence that have previously been used to motivate a normative framework, but this does not mean that the cases I am considering here are the only ones that can be used in support of the normative approach. Secondly, the discussion will provide a fairly cursory overview of the evidence in favour of a normative framework. In the next Chapter, I will focus in on the merits of the approach in much greater detail by examining a normative framework in the context of an interventionist theory of causation. For now, I want to offer up some very general motivations philosophers have given in defence of the view. Finally, within each sub–section I'll highlight what kind of normative considerations are being invoked to satisfy the desiderata. Specifically, I'll stress that all kinds of norms, including statistical norms and the vast array of prescriptive norms get put to work in explaining the desiderata; no set of norms is privileged over any other. This point crucial to the discussion that takes place towards the end of the Chapter where I argue

that a non-hybrid form of normativism is a more successful theory of causation compared to a hybrid form.

2.4 Desiderata associated with Causal Theories

2.4.1 Ordinary Causal Judgements

A recent flurry of empirical evidence has suggested that normative considerations heavily influence our causal judgements. The focus of much of this research has been on a phenomenon that has become known as the 'Knobe effect' (as much of the work on this has been done by Joshua Knobe). Put simply, this is the effect normative considerations have on the extent to which we agree with a causal statement. The most influential case where this effect has been shown to occur is Joshua Knobe and Ben Fraser's (2008) pen vignette. Knobe and Fraser presented their participants with a short vignette which is as follows:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own.
>
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mails them reminders that only administrators are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message… but she has a problem. There are no pens left on her desk. (2008, pp. 143-144)

Having been presented with the vignette, the participants were then asked whether they agreed or disagreed with the following statements:

1. Professor Smith caused the problem.
2. The administrative assistant caused the problem

Participants were much more likely to agree with (1) and disagree with (2). Knobe and Fraser argue that the results can be explained by the fact that it is normal for the administrative assistant to take pens, but it is not normal for faculty to do the same. This suggest that whether people's judgements about causation are sensitive to considerations of what is normal for a given context. These results are easily explained by a normative account of causation. This study was soon followed by similar experiments which demonstrate the influence of norms on causal judgements. When two agents login to a computer (Knobe & Fraser 2008) or enter a room with a motion detector (Icard et al., 2017, Kominsky et al., 2015), but only one agent is allowed to perform this action, the norm-violating agent is judged to a cause (or more of a cause) for the subsequent consequences of their actions.

It's worth noting that Hitchcock and Knobe (2009) were able to replicate the results of these findings in cases which did not involve agency:

> A machine is set up in such a way that it will short circuit if both the black wire and the red wire touch the battery at the same time. The machine will not short circuit if just one of these wires touches the battery. The black wire is designated as the one that is supposed to touch the battery, while the red wire is supposed to remain in some other part of the machine.
>
> One day, the black wire and the red wire both end up touching the battery at the same time. There is a short circuit. (p. 604)

In this case, if the black wire hadn't touched the battery, there wouldn't have been a short circuit; and if the red wire hadn't touched the battery, there wouldn't have been a short circuit. Even so, participants tended to agree that the fact that the red wire touched the battery caused the machine to short circuit and disagree with the statement that the fact that the black wire touched the battery caused the machine to short circuit. These judgements can be explained by the fact that the red wire violates a design norm or functional norm when it touches the battery, and therefore, we choose the red over the black as a cause of the short circuit.

It's important to emphasise that for Hitchcock and Knobe the types of norms that make up the relevant conception of normality and abnormality are not derived from any specific category

of norm. Rather they understand normality to be a kind of heuristic comprised by a wide variety of statistical and prescriptive categories:

> [I]t might be thought that we need to say which specific type of norms end up playing a role in the way people assess the relevance of counterfactuals. Yet although there clearly are differences between these kinds of norms, it also seems that it is not merely a sort of pun that they have come to be denoted by the same word. There really does seems to be some important way in which these different kinds of norms are connected in our ordinary way of understanding the world. Hence, when we say of a person that she is 'abnormal,' we do not typically have in mind either a purely statistical judgement or a purely prescriptive one. Instead, we seem to be making a single overall judgement that takes both statistical and prescriptive considerations into account. (2009, p.589)

Research into the Knobe effect and related phenomena, has demonstrated how our causal judgements are sensitive to norm violations. Most of this research has focused on how the normality of an agent's action affects that agent's own causality, not anyone else's. However, a further study conducted by Kominsky et al. (2015, 2017) has shown that the extent to which one agent is perceived to have caused an outcome may be affected not only by the normative status of her own actions, but also the normative status of other people's actions. Kominsky et al. (2017) call this phenomenon 'causal superseding'. They describe it as follows: "suppose an outcome depends on a causal factor C as well as an alternative causal factor A, such that the outcome will only occur if both C and A occur. Then people will be less inclined to say that C caused the outcome if A is abnormal than if A is normal" (p. 81). I will not outline the details of the experiments here, but it's important to note that Kominsky et al. (2015, p. 208) argue that the results of the experiments show that all categories of norms inform our judgements about causal superseding. They emphasise, for instance, that not only do social and moral norms play a significant role in our causal attributions, so do pragmatic and statistical norms.

Before moving onto the second desiderata, I want to clarify in what sense empirical evidence is supposed to serve as evidence in favour of a normative approach. Proponents of normativism, at least as I've characterised them, are concerned with arguing that causal *facts* are sensitive to normative considerations. The results of the experiments, however, seemingly provide reasons in favour of believing that our causal *judgements* are sensitive to normative considerations. So,

in what sense does the empirical evidence give us reasons to think that causal facts are sensitive to norms? The proponent of the normative approach is trying to offer the best explanation of our intuitions about the acceptability of causal judgements in these cases by claiming that these intuitions reflect the real truth-values of those statements. This is to say that what explains the fact that our causal judgements are sensitive to normative considerations is the fact that causation itself is normative. The approach is 'bottom-up' rather than 'top-down'.

2.4.2 Distinguishing between Background Conditions and Causes

For any given effect, there will be a considerable number of candidate causes that the event counterfactually depends upon. Yet when asked to identify the event that caused the effect in question, we tend to respond with one or two events that we select from the almost endless set of candidates. The remaining members of this set are hardly ever mentioned, and when they are mentioned, they are seen to be merely background conditions for the occurrence of the effect. For instance, if a forest fire occurs, the forest ranger will likely select the lightning strike as a cause of the fire and relegate the presence of oxygen to a background condition, despite the fact that the presence of oxygen is necessary for the occurrence of the fire.

Adopting a normative framework allows us to make sense of our selections. The thought goes that the difference between background conditions and actual causes depends upon whether the event in question represents a normal or deviant state of affairs. Things which represent a normal state of affairs are background conditions, whilst things which deviate from what's normal are actual causes. The presence of oxygen in our atmosphere is a normal state, and therefore, its presence is relegated to a background condition, whereas lightning strikes are statistically very unusual, thus they are promoted to actual causes.

I suspect that the types of norms at play when it comes to discriminating between causes and background conditions will in the majority of cases be statistical norms. In the forest fire case, for instance, what makes the presence of oxygen normal is presumably to do with the frequency with which oxygen is found in the earth's atmosphere. There's no sense, as far as I can see, that oxygen is prescriptively supposed to be in the earth's atmosphere. With that said, we can imagine examples where a state of affairs is normal because prescriptive norms make it so. For instance, you might live in a neighbourhood where it's normal for you to maintain an

immaculate front lawn. What makes the immaculate front lawn a normal state of affairs is the fact that it aligns with a social norm particular to your neighbourhood. There might also be a sense in which having an immaculate front lawn would be statistically normal in your neighbourhood, but the statistical norm seems to be the result of the fact that having an immaculate front lawn is first and foremost a social norm; if one lives in that neighbourhood, one *ought* to be seen with a manicured lawn. This is all to say that the types of normative considerations that inform what makes a state of affairs normal or abnormal, and consequently determine the difference between background conditions and actual causes, will be derived from statistical norms and sometimes prescriptive norms.

2.4.3 Causation by Omission

One standing problem in treatments of causation arises with omissions. We have already touched upon the problem earlier in the Chapter. Suppose that my friend promises to feed my fish whilst I'm away, but fails to do so, and the fish dies. People tend to blame the friend and speak of her omission to feed the fish as a cause of its death. But what is so special about my friend? My neighbour, my gran, Bono and indeed everyone else also omitted to feed the fish. Indeed, metaphysically speaking, my friend and Bono seem perfectly alike: neither actually fed the fish, and each is such that, counterfactually, if she/he had fed the fish then the fish would have survived.

Is there a way to capture the judgement that my friend is a cause of the fish's death, along with the judgement that Bono is not a cause? It's worth noting that some believe that these judgements are not worth capturing. For instance, Helen Beebee (2004) and Achille Varzi (2007) claim that causation by absence is not causation at all but rather causal explanation. They argue that absences can figure in causal explanations of events, but not in virtue of being bone fide causes or effects. Rather, claims about absences provide us with causal information. Phil Dowe (2000) offers an alternative explanation and claims that absence causation is merely quasi-causation, a closely related but decidedly non-causal concept. According to Dowe, absences can quasi cause something to occur, but they cannot actually enter into causal relations. For Dowe, quasi-causation is possible causation, not actual causation. Even so, intuition strongly suggests that absences do enter into causal relations. So let us suppose that we want to capture a distinction between my friend and Bono. How might one do so?

One popular strategy, which is now becoming orthodoxy, is to appeal to the normative status of the absence. Tony Honoré (1999), Sarah McGrath (2005), and Peter Menzies (2007) for instance, argue that the absences which are causally efficacious are those that deviate from the normal course of events. Causally inert absences, on the other hand, are the ones that do not deviate from the normal course of events. According to this strategy, my friend is a cause of the fish's death in virtue of the fact that my friend's failure to feed the fish deviates from what would normally occur, inasmuch as it is normal for friends to keep their promises. By contrast, Bono's omission is not a cause since his failure to feed the fish does not represent a deviation from a norm, on the contrary, his omission is quite ordinary.

As with the preceding desiderata, it's important to clarify that the types of norms invoked to meet this desideratum come from a wide range of categories. Indeed, in her account of absence causation, McGrath (2005) makes a great and conscious effort to emphasise that the sense of normal being used to determine the normative status of an absence is a very inclusive one:

> The notion of the normal I have in mind is highly abstract and applies very generally: to actions, the behaviours or artifacts, and the behaviours of both biological and non-living systems. […] It is normal for x to φ iff x is *supposed* to φ. People are supposed to keep their promises (it is normal for them to keep their promises); alarm clocks are supposed to ring at the set time (it is normal for them to ring at the set time); hearts are supposed to pump blood (it is normal for them to pump blood); the rain is supposed to come in April (it is normal for it to come in April); water is supposed to flow downhill (is it normal for it to flow downhill). (2005, p. 138. Original emphasis)

For McGrath, something is normal to the extent that it is supposed to happen, and as her examples illustrate, the relevant sense of 'supposed to' is intended to be read liberally. The 'supposed to' that applies to keeping promises is presumably one that derives from a moral norm, the 'supposed to' to that applies to alarm clocks ringing and hearts pumping blood derives from a functional norm. In contrast, the 'supposed to' that applies to rain falling in April and water flowing downhill is perhaps a statistical norm, or especially in the latter case, a norm derived from physical laws.

2.5 On the Incompatibility between Causal Realism and Normativism

Given the purported benefits of normative views, one might well wonder whether we can marry the approach with causal realism. Might there be a way for the realist to incorporate normative considerations into her account without compromising her realist credentials? I mentioned in the Introduction that contemporary discourse has answered in the negative. Many believe that if one appeals to normative considerations in order to determine the causal facts, then one undermines the metaphysical assumptions central to the realist position.

Sara Bernstein articulates the perceived incompatibility as follows:

> Metaphysical realists hold that causation is a joint-carving, mind-independent relation: something 'in the world' that would exist exactly the way it does regardless of human thoughts or interventions. But views of causation that incorporate human thought and agency are taken to threaten the mind-independence and objectivity of the causal relation. (2017a, p. 217)

And Georgie Statham writes:

> Certainly the claim that judgements of actual causation are influenced by normative commitments suggests that this concept isn't directly connected to fundamental metaphysics — prescriptive norms like the requirement that we keep promises don't seem to be the kinds of things that 'carve nature at the joints' as metaphysicians like to put it. (2017, p. 458).

Furthermore, when discussing absence causation, self-proclaimed causal realist, Helen Beebee, says:

> Take the violation-of-norms part of the definition. If we take the definition to give the *truth conditions* of causation by absence claims, then causal facts about absences depend in part on normative facts: facts about whether a moral or epistemic or other norm has been violated. But nobody within the tradition of the metaphysics of causation that I'm concerned with here thinks that causal facts depend on human-dependent norms. (2004, p. 297. Original emphasis.)

Here and elsewhere in the literature, the specific concern appears to be that the realist cannot appeal to normative considerations because normative considerations are mind-*dependent*. Notions like normality and abnormality are thought to be comprised by how humans experience and interact with the world. Insofar as norms are mind-dependent the realist is not entitled to use them to determine causal facts since doing so would render the causal facts themselves mind-dependent. This cuts against the realist claim that facts about causation exist independently of us. Put simply, the status of causation under a normative framework is incompatible with what the realist takes it to be.

A mind-dependent notion of causation generates further, more specific concerns for the realist. Halpern and Hitchcock (2015) summarise the concerns as follows: "the worry is that the incorporation of norms will render causation: (i) subjective, (ii) socially constructed, (iii) value-laden, (iv) context-dependent, and (v) vague" (p. 431). Causation would be subjective to the extent that the truth conditions for causal facts depend upon subjective perceptions of what's normal or abnormal. Causation would be socially constructed insofar as what's normal can be a product of social norms. Further, since some norms incorporate values causation would become value-laden. For instance, the idea that one ought to keep one's promises expresses the thought that it is in some sense valuable or good to keep one's promises. Hence, if we let norms about promise keeping determine causal facts, then causal facts become entangled with values. Causation would also become context-dependent since norms typically come in and out of being depending on the context — it might be normal for Tom to wear a three-piece suit to a wedding, but abnormal for him to wear it to a football match. And since normality is something that admits of degrees — something can be more or less normal — it can be unclear whether some state of affairs are normal simpliciter. Hence, appealing to normality would render causation vague. Causation has none of these features says the realist.

These reasons have led many to endorse the following:

INCOMPATIBILITY: causal realism is incompatible with a normative approach to causation.

In what follows, I argue against INCOMPATIBILITY. Doing so, however, requires getting more precise about the argument offered up in support of INCOMPATIBILITY. In particular, we need to disambiguate what exactly is entailed when we say something is 'mind-

independent', and in particular, what kind of notion the realist is invoking when she describes causation as a 'mind-independent' relation.

2.5.1 Clarifying INCOMPATIBILITY

There are at least two ways something could be mind-independent. Firstly, one could be claiming that domain X is mind-independent if and only if the truths and falsehoods about X are not *determined* by any mind or perspective. Common sense suggests that mathematical facts are of this kind. The claim that the area of a circle is equal to πr^2 is not a claim that is made true in virtue of any mind; the formula is true regardless of anything humans have think or say about it.[9] Other paradigmatically mind-independent properties include things like shape and size. Mount Everest has a height of 8,849 meters, and although it may *appear* to us to be taller or shorter depending on where one is viewing it from, its height will remain 8.849 meters because this property is said to be determined independently of our perception of it. To clarify this first notion of mind-independency further, consider how Russ Shafer-Landau describes how the concept operates in the moral domain:

> Realists believe that there are moral truths that [hold] independently of any preferred perspective, in the sense that *the moral standards that fix the moral facts are not made true by virtue of their ratification from within any given actual or hypothetical perspective*. That a person takes a particular attitude toward a putative moral standard is not what makes that standard correct. (2003, p.15. Original emphasis). [10]

In contrast to things like mathematics, mass, height, and shape, canonical categories of mind-*dependent* phenomena include things like beauty and aesthetic taste. If beauty is, as it were, in the eye of the beholder, then whether X is beautiful is not determined by anything that exists outside of the mind. On the contrary, whether X is beautiful is determined from the point of view of a particular perspective.

---

[9] To be more precise, the ratio of a circle's area to its radius is mind-independent, but this notation – πr^2 is mind-dependent.

[10] Shafer-Landau prefers the term 'stance-independent' to mind-independent due to the latter's ambiguity and obscurity.

So that is the first notion of what it could mean to be mind-independent. The second notion of mind-independence I will refer to as mind-independence*. A domain X is mind-independent* if and only if the truths and falsehoods about X *obtain* in the absence of minds. That is, if we suddenly ceased to exist this would not make it so that facts about X would suddenly cease to exist. Again, common sense suggests that mathematical facts are of this sort. If humans were to suddenly zap out of existence, the area of a circle would continue to be πr^2. Similarly, if humans suddenly ceased to exist certain properties like an object's shape and size would presumably continue to obtain — Mount Everest would continue to stand at 8,849 meters high, even if there were no minds around to perceive the mountain's height. In contrast to these categories of things, certain claims about mental states are often taken to express paradigmatic examples of mind-*dependent** phenomena. For instance, claims like 'A is experiencing pain' can only be true when there is a mind in the world to entertain the phenomenological experience of pain. Pain exists insofar as one has an experience of it, so facts about pain can only obtain when there is a mind to experience painful states.

To clarify these two notions further, let's look to the moral domain again. Moral Relativists claim that various conventions and social agreements are the things out of which morality is constructed. Moral facts are therefore constructed from the attitudes, outlooks, actions and activities undertaken from a particular standpoint. If there were no individuals to make agreements or establish conventions, then there would be no moral reality. So understood, relativism construes morality as mind-dependent and mind-dependent*. In contrast, Plato's realism implies that moral facts are out there in the world independently of our talking and thinking about them, and further that these facts are not reducible or depend upon other non-moral facts, properties or objects like minds and observers. So understood, Plato's realism sees moral reality as mind-independent and mind-independent*. Furthermore, it's possible for a theory to imply one of these notions of mind-independency without also implying the other. For instance, utilitarianism takes facts about right and wrong to be independent of any particular perspective; right acts are those which increase utility, this is true regardless of minds, speakers or observers. Still for the utilitarian, right and wrong are only ever instantiated when there are minds in the world to experience pleasure and pain. Hence, utilitarianism is mind-independent and mind-dependent*.

From these formulations, we can draw the below chart. An 'x' denotes that the theory has the relevant property.

	Mind-Independence (truths and falsehoods about X are not *determined* by any mind or perspective.)	Mind-Independence * (truths and falsehoods about X *obtain* in the absence of minds or perspectives.)
Moral Relativism		
Platonic Realism	x	x
Utilitarianism	x	

(Table 1 in Chapter 2)

I take the causal realists to be arguing that causal facts are mind-independent and mind-independent*. This is to say that they claim (i) causal facts are not *determined* by any mind or perspective and (ii) causal facts *obtain* in the absence of any minds or perspectives.[11] Claims such as: 'the rainfall caused the puddles on the ground' are not made true or false in virtue of any mind nor will they obtain or fail to obtain in the absence of minds. Let's apply this clarification to the thesis under discussion. When philosophers argue that realism is incompatible with a normative approach on the grounds that appealing to norms to determine causal facts will undermine the realist assumption that causation is a mind-independent feature of the world, what they mean by 'mind-independence' is, we can suppose, both notions of mind-independent and mind-independent *.

2.6 Challenging INCOMPATIBILITY

Having made some headway into clarifying the argument, I'm now in a position to show why the argument does not entail an incompatibility between realism and an appeal to normative considerations. The most straightforward strategy is to challenge the claim that normative considerations are mind-dependent. If one can show that facts about normality and abnormality are mind-independent/mind-independent*, then the realist would be entitled to appeal to such

[11] Not all causal facts will be mind-independent* for the realist, because some will make reference to minds. For example, 'the hot coffee caused Ali pain' is a mind-dependent* claim. Nonetheless, I understand realism as arguing that there are at least some causal facts in the absence of minds.

facts. After all there would be mind-independency ‘all the way down’. This is the strategy I will take up here. Let’s begin by considering statistical norms.

A statistical norm represents the frequency with which a certain event occurs or the frequency with which a state of affairs obtains. Statistical norms can manifest themselves across a huge range of domains, and some of these domains will be mind-dependent/mind-dependent*. For example, norms that encode statistical information about mental states or mental processes will be mind-dependent* since the information will depend upon the existence of those minds. Nevertheless, it seems to me that barring these sorts of cases, the majority of statistical norms will be mind-independent and mind-independent*. They are mind-independent insofar as whether something is statistically normal is not determined by any particular mind or perspective, and they are mind-independent* in the sense that they obtain in the absence of any mind or perspective. Take as an example an everyday statistical norm such as ‘Glasgow has more winter rainfall than Milan’. This norm is not made true by any mind, nor would it cease to obtain if minds were suddenly zapped out of existence.

If true — in particular, if it’s true that certain statistical norms are mind-independent/mind-independent* — then the realist could appeal to these norms when determining the causal facts without undermining their metaphysical commitments. To illustrate how this might work, suppose that one winter the residents of Glasgow report more respiratory ailments than is typical for that time of year. And suppose that the very same winter the city had less rainfall than it would usually have. The realist can claim that the increase in respiratory ailments was caused by Glasgow’s dry winter, and importantly, it would be consistent for them to claim that what *makes* the weather a cause, as opposed to some other event, is partly due to the fact that the weather represents a deviation from what is statistically normal.

It's worth noting that if some statistical norms are mind-independent/mind-independent*, then this means that some expectational norms will also be mind-independent/mind-independent*. For recall that expectational norms are generated by what is statistically likely to occur in the future. If Glasgow has more rainfall than Milan in winter, there’s a sense in which Glasgow *ought* to have more rain than Milan this coming winter. This ‘ought’ is entailed by a statistical (rather than prescriptive) understanding of normality. So, if expectational norms can simply be reduced to statistical norms, then presumably they possess the same mind-independent/mind-

independent* character as statistical norms, and therefore can be deployed by the realist to determine causal facts.

At this point in the discussion, I'll pause to note an interesting implication about what's just been said about statistical and expectational norms. I've shown that one can appeal to these categories of normative considerations whilst consistently holding that causation is mind-independent and mind-independent*. Interestingly, by establishing that some causal facts are mind-independent* (i.e. obtain in the absence of minds) we can make the more general claim that causation *simpliciter* is mind-independent*. Why might this be? In order to establish that causation simpliciter is mind-independent* one only needs to demonstrate that causal facts exist in the absence of minds, and this is precisely what has been done through the discussion of statistical norms. We've seen that one can appeal to statistical norms to determine causal facts, and that this is consistent with the thought that such facts would exist independently of us. By showing that at least *some* causal facts exist mind-independently* under a normative framework, we can make the more general (and significant) claim that causation simpliciter is mind-independent* under such a framework. Going forward then, there is no need to consider whether a category of norms is mind-independent*, and so I will only consider whether the category of norms is mind-independent (i.e. whether the norms are determined by minds and perspectives).

Next let us move to consider prescriptive norms. First take the category of moral norms. It seems like whether causal realism is compatible with an appeal to moral considerations will depend upon which moral theory is correct, because, as we have seen, different moral theories deliver different verdicts about the mind-independency of moral facts. If moral relativism is correct then moral facts are determined by our collective attitudes, outlooks, actions and activities undertaken from particular standpoints. So understood, relativism construes morality as mind-dependent. The causal realist, therefore, would not be entitled to appeal to moral norms to determine the truth or falsity of causal claims whilst maintaining that causation is mind-independent. By contrast, both Platonic realism and Naturalist Realists (such as some forms Utilitarianism) take the truth value of moral facts to be determined independently of any minds or perspective, thus, if either of these theories were correct, then an appeal to moral norms to determine causal facts would not undermine the realist's metaphysical commitments. Consequently, whether a causal realist position is compatible with an appeal to moral norms depends upon the nature of moral facts themselves.

Things are more clear-cut when it comes to other prescriptive norms such as social norms. Social norms are rules that govern acceptable behaviour, beliefs and attitudes in a particular social group or culture. They are context-sensitive and contingent upon implicit or explicit societal agreements and social practises. For this reason, social norms are most obviously mind-dependent; whether they are true depends upon the particular perspectives of groups of individuals. In South Africa, for instance, it is customary for funeral attendees to wear the colour red, and what makes this social norm true is the fact that a collective group of agents — namely, South Africans — share the same belief that they manifest via social practise. Shift perspectives and practises, and we find this particular social norm to be false (in China it is socially unacceptable for mourners to wear red). In terms of causation, this means that the causal realist cannot invoke social norms in order to determine causal facts, since doing so would render the causal facts themselves mind-dependent.

Although the preceding discussion includes only a brief survey of some canonical norms, it's enough to demonstrate that the dominant argument given in support of INCOMPATIBILITY fails. Of the relatively small list of norms I evaluated, the realist could appeal to certain statistical norms, expectational norms, and depending on the correct moral theory, moral norms, whilst preserving the thought that causation is a mind-independent/mind-independent* natural relation.[12]

Now there is a way to resist this line of thinking. One could argue that although there are certain sets of realist-friendly norms, the realist is not entitled to cherry pick only these norms to determine causal facts, because doing so would be objectionably arbitrary. As I stated earlier, the strongest justification for a normative approach is its success in satisfying a vast range of seemingly disparate desiderata associated with theories of causation. To recap the desiderata included: the ability to distinguish between background conditions and causes, the ability to identify which omissions participate in causal relations, and the ability to produce causal judgements that fit the folk conception of causation. When outlining these desiderata, I argued that all types of norms equally support the fulfilment of these desiderata. That is, every category

[12] Due to scope constraints, I could not analyse each category of normative consideration to determine whether it is mind-independent/mind-independent*. Nonetheless, I suspect there will be even more norms than the ones I list here that a realist can appeal to such as certain functional norms concerning biology. For example, the function of a plant's chloroplast is to conduct photosynthesis, this biological norm is not determined by any mind, nor will it fail to obtain in the absence of such minds.

of norm including moral norms, norms of etiquette, cultural norms, epistemic norms, statistical norms, functional norms, and so forth, function equally well in their ability to meet each objective. Given that they all fulfil this function equally well, one might argue that there's no non-arbitrary way to discriminate between the types of norms one can employ into one's causal theory. To put it another way, in terms of acting as truth-makers for causal claims, there just isn't any feature that some norms have and others lack in virtue of which some norms can be appealed to and others cannot. Therefore, the realist is not entitled to add, say, statistical norms into a causal theory without also adding social norms. So, to the extent that incorporating one set of normative considerations requires incorporating all sets of normative considerations, the causal realist cannot appeal to normative considerations simpliciter, making INCOMPATIBILITY true.

There is an avenue of reply available to the realist when faced with this response. Whilst the realist can accept the claim that each category of norms functions equally well in its ability to satisfy the desiderata, they can deny that there's no non-arbitrary way to discriminate between the kinds of norms one can employ in a causal theory. Specifically, they can argue that the realist is justified in making a distinction between norms, and further that this distinction can be cut in a way that distinguishes between the set of norms which do undermine a realist conception of causation from those that do not. This response turns on the thought that in addition to the three desiderata I have named, there is another unarticulated desideratum associated with theories of causation; namely, that theories of causation ought to deliver causal verdicts which are compatible with a mind-independent and mind-independent* concept of causation. We might call this 'the mind-independency(*) desideratum'. In terms of incorporating normative considerations, the mind-independency(*) desideratum effectively demands that we draw a distinction between the norms which are realist-friendly from those which are non-realist-friendly, thereby offering a seemingly non-arbitrary way to distinguish between types of normative considerations. One can then invoke certain categories of normative considerations to determine causal relations whilst retaining realism about causation, thus we're back to the conclusion that INCOMPATIBILITY is false.

2.7 An Argument in Favour of Unrestricted Normativism

Thus far I have demonstrated that there are ways for the causal realist to resist INCOMPATIBILITY. I have argued that causal realism is not necessarily incompatible with a

normative framework, and that realism can appeal to a restricted set of normative considerations to determine causal relations. These normative considerations include statistical norms, expectational norms, and potentially, moral norms. This conclusion has significant implications. For one thing it opens the door to developing a 'hybrid' view of causation that blends realist metaphysics with a restricted normative framework. We might call this kind of hybrid view "restricted normativism".

Restricted normativism preserves the realist conception of causation in virtue of drawing on those normative considerations that do not violate the mind-independency(*) desideratum. But not all causal theorists who want to appeal to norms will accept mind-independency(*) as a desideratum. These folk will be happy to deny that theories of causation ought to deliver causal verdicts which are mind-independent and mind-independent*. The denial of the mind-independency(*) desideratum leads to a non-hybrid form of normativism, one that is not restricted in the kind of normative considerations it can draw upon. This view, which we might call "unrestricted normativism", is entitled to appeal to all categories of norms including those which are mind-dependent. Most contemporary normative views will fall under this umbrella inasmuch as they do not discriminate between the kinds of normative considerations that determine causal facts. Unrestricted normativism is evidently incompatible with a realist conception of causation.

Alongside unrestricted and restricted normativism, we also have what might be termed a "pure realist" view left on the table. The pure realist view conceives of causation as a mind-independent feature of the world and it does not appeal to normative considerations at all to determine causal relations. It's fair to say that traditionally philosophers working on causation have followed a pure realist approach. Lewis's (1973) counterfactual analysis and Mackie's (1980) regularity account, for instance, conceive of causal relations as mind-independent and do not invoke any kind of normative considerations in order to determine them (at least not explicitly anyway).

Before concluding, I will compare the newly outlined restricted normativism with unrestricted normativism. I set aside discussion of pure realism because I want to compare causal views that incorporate a normative framework. In what follows, I will provide some reasons in favour of thinking that unrestricted normativism is a more successful theory of causation than restricted normativism. My argument turns on the claim that restricted normativism cannot

adequately satisfy the other desiderata discussed in Section 4. Whereas, unrestricted normativism can adequately satisfy the desiderata, and furthermore it can do so in a way that is intuitive and unified. Insofar as unrestrictive normativism is better able to explain a plethora of causal concepts, I suggest that it is more successful.

The first step in the argument is to point out, as was done previously, that all kind of norms function equally well in their ability to meet the various desiderata: distinguishing between background conditions and causes, identifying which omissions participate in causal relations and producing causal judgements that fit the folk conception of causation. That being the case, a view which meets mind-independency(*) by restricting itself to only certain kinds of norms will do worse in meeting these three desiderata than a view which is unrestricted in the kinds of norms it can appeal to. This is to say that mind-independency(*) is at odds with the other desiderata. Hence a theory which aims to satisfy mind-independency(*) will do so at the expense of being able to meet the other objectives associated with theories of causation.

To illustrate one way in which unrestricted normativism is better able to satisfy the desideria, consider a cause of omissive causation. In particular, consider a case of omissive causation where the causal relations appear to be determined by the violation of a social norm. Imagine that McEnroe refuses to shake the hand of his opponent after an intense tennis match, his opponent then throws his arms up in the air in disbelief. It's natural to think that the opponent's reaction was caused by the fact that McEnroe omitted to shake his hand. Unrestricted normativism can make sense of this judgement because it is entitled to invoke social norms to determine causal relations. It can point to the fact that failing to shake a competitor's hand after a tennis match violates a social norm, making McEnroe's failure a cause of his opponent's reaction. However, advocates of restricted normativism will have a difficult time establishing this causal relation. Given that social norms are mind-dependent, restricted normativists are not entitled to invoke social norms as a means to determine that it was McEnroe that caused his opponent to react the way he did.[13]

[13] There might be another way to explain why McEnroe's omission is a cause without appealing to *social* norms. One could argue that it is statistically normal for competitors to shake hands after a competitive game. Hence, McEnroe's omission is a cause of his opponent's reaction in virtue of the fact that his omission violates a statistical norm. I've argued that statistical norms are mind-independent/mind-independent*, so this kind of explanation would be able to satisfy the desideratum about omissions, whilst also adhering to the mind-independency desideratum. However, I think appealing to only statistical norms doesn't tell the whole story. It seems like the existence of the statistical norm is *dependent upon* the existence of the social norm; shaking an opponent's hand after

There are many more instances where unrestricted normativism has the upper hand over restricted normativism in terms of meeting the desiderata. As a further illustration, suppose this time we want to distinguish between background conditions and causes. Suppose Hollie is trying to park her car in a busy supermarket carpark. After doing laps of the carpark she finds a spot, and parks her car. But in her rush Hollie forgets to buy a parking permit, and on her return, she notices a parking ticket on her windshield. Presumably, we want to say that Hollie's failure to purchase a parking permit caused her to receive a parking ticket. Perhaps we would also want to say that Hollie parking the car where and when she did caused the parking ticket. But presumably we would not want to say that Hollie's getting a parking ticket was caused by all the other vehicles in the carpark being parked within the white lines. This is surely a mere background condition to Hollie being issued a parking ticket. Unrestricted normativism can accommodate for this distinction. It can say that parking within the white lines is socially normal behaviour, thus it should be relegated to a background condition. Restricted normativism, on the other hand, cannot appeal to social norms to draw this distinction. Indeed, several canonical realist theories of causation which do not appeal to social norms, such as the simple counterfactual analysis, will conclude that the other motorists parking within the white lines do cause of Hollie's parking ticket, since if they had not parked so efficiently Hollie would not have been able to park her car in the first place.

Furthermore, unrestricted normativism is more effective at producing causal verdicts that fit the folk conception of causation. This argument find support in the work of Hitchcock and Knobe (2009). Recall that for them, the empirical evidence demonstrates that our ordinary causal judgements are sensitive to judgements about what's normal and abnormal, which are themselves understood to be single, overall heuristic evaluations that are comprised of the diverse range of prescriptive and statistical norms. This suggests that our causal judgements do not privilege any particular set of norms over any other, nor do they privilege the wider category of statistical norms over prescriptive norms or vice versa. To leave out some norms from one's framework then, would be to inadequately capture our ordinary judgements about what's normal and abnormal. Consequently, restricted normativism would be less successful at accounting for the folk causal judgements.

---

a competitive game is first and foremost an example of social convention, and the existence of any statistical pattern seems to fall out as a result of this convention.

These examples demonstrate that unrestricted normativism can meet several desiderata associated with theories of causation in a way that restricted normativism cannot. Furthermore, unrestricted normativism is able to satisfy these seeming disparate desiderata in a cohesive and comprehensive manner by invoking a unified normative framework. The view therefore possesses significant theoretical virtues, and it seems to me that these virtues are not outweighed by the fact that in securing them we must jettison the intuition that causation is a mind-independent feature of the world, even if this intuition is widely embedded into our causal theorising. My conclusion then, is that a full blown, unqualified normativism is a more successful theory of causation than a normative view that is constrained by realist metaphysics.

2.8 Conclusion

I began this Chapter by delineating two meta-causal views of causation — normativism and realism. I said that contemporary philosophical discourse conceives of these two meta-causal views as incompatible approaches about causation. To put it simply, the idea was that normativity is determined by minds and perspective, hence normative considerations are understood to be mind-dependent. The causal realist, therefore, is not entitled to deploy normative notions to determine causal connections, lest those connections themselves be rendered mind-dependent. This narrative supported an argument I termed INCOMPATIBILITY:

INCOMPATIBILITY: causal realism is incompatible with a normative approach to causation.

It has been the aim of this Chapter to push back against the prevailing discourse by arguing against INCOMPATIBILITY. To do this, I first clarified the arguments given in support of the claim. Specifically, I disambiguated what exactly is entailed when we say something is ‘mind-independent’, and then made explicit in what sense the realist takes causation to be a ‘mind-independent’ relation. After making these clarifications, I then went on to show that some normative considerations are mind-independent in the sense the realist conceives of causation as being mind-independent. Thus, I argued that the realist is entitled to appeal to these mind-independent normative considerations to determine the truth value of causal claims without compromising her metaphysical commitments.

This discussion left three approaches to causation on the table. The first is a pure realist view which conceives of causation as a mind-independent natural relation in the world and does not appeal to any normative considerations in order to determine the causal facts. The second is a hybrid view comprising of realism and normativism. I referred to this view as a restricted normativism since it can only draw on a limited set of normative considerations; namely those that are mind-independent/mind-independent*. The third kind of view, which I termed unrestricted normativism, does not aspire to preserve a realist conception of causation, and is therefore justified in incorporating all categories of norms to determine causal facts. In the final part of this Chapter, I argued that unrestricted normativism can adequately and uniformly satisfy several seemingly disparate desiderata associated with theories of causation in a way that restricted normativism cannot. Insofar as this is the case, I concluded that unrestricted normativism promises a more success approach to causation.

CHAPTER 3

Interventionism, Causal Realism and Normative Considerations

3.1 Introduction

In the previous Chapter I introduced two views about causation:

1) Causal realism: causal facts are mind-independent
2) Normativism: causal facts are partly determined by normative considerations

I said that contemporary philosophical discourse conceives of these views as incompatible approaches to causation. Put simply, the idea was that what's normative is determined by minds and perspectives, hence normative considerations are presumed to be mind-dependent. The causal realist, therefore, is not entitled to deploy normative notions to determine what causal connections there are in the world, lest those connections themselves be rendered mind-dependent. I referred to this argument as INCOMPATIBILITY. I argued, contra to the prevailing narrative, that INCOMPATIBILITY is false because some normative considerations are mind-independent, thus the realist would be justified in appealing to them when determining causal relations without undermining her metaphysical commitments. This argument left us with three meta-causal views on the table. Firstly, a "pure realist" view which conceives of causation as a mind-independent natural relation and does not appeal to any normative considerations in order to determine the causal facts. Secondly, a hybrid view comprising of realism and normativism. I termed this view "restricted normativism" since it can only invoke a limited set of normative considerations, specifically those that are mind-independent/mind-independent*. Finally, a view I called "unrestricted normativism", which is not motivated to uphold a realist conception of causation, and thereby is entitled to draw upon all categories of norms including those that are mind-dependent.

In this Chapter, I'm going to analyse where one particular account of causation sits in regards to the divide between these three meta-causal views. The account I have in mind is James Woodward's (2003) interventionism. Roughly, interventionism says that *X* is a cause of *Y* where there exists an intervention or manipulation of *X* that would bring about an associated change in *Y*.

I have several motivations for focusing on Woodward's interventionism. Firstly, interventionism has enjoyed a huge surge in popularity over the last couple of decades. To present only a few paradigmatic examples of its success and fruitfulness in areas of philosophy, interventionism has been applied to: mechanistic explanation (Glennan 2002), the philosophy of biology (Waters 2007), mental causation (Campbell 2007), (Shapiro and Sober 2007), laws and explanation (Hitchcock and Woodward 2003), (Leuridan 2010). Its popularity has led Alexander Reutlinger (2013) to label interventionism "the new orthodoxy" (p. 6) in these philosophical domains. Not only is the view permeating through wide areas of philosophical discourse, it is also gaining serious traction in computer science, social science, artificial intelligence, economics, and statistics. Woodward's 2003 delineation of the theory is the most influential philosophical account of interventionism, it therefore makes sense for me to focus on his exposition of the view.

The second reason for focusing on Woodward's interventionism derives from the fact there's a small but bourgeoning debate as to whether his theory is compatible with causal realism. This is chiefly because Woodward's view makes use of causal models, and there is some controversy around whether such causal models need to rely on notions of normality and abnormality for their success. For those who believe that one can generate correct causal models without invoking notions of normality at all, like Woodward himself, can, other things being equal, maintain that interventionism is compatible with a realist conception of causation. According to the taxonomy I produced in the previous Chapter, this position would exemplify a pure realist approach to causation. A second avenue open to the interventionist is to concede that correct causal models need to incorporate notions of normality, but to hold that the relevant notions of normality deal in only mind-independent norms. This would place interventionism under the umbrella of restricted normativism, again making it compatible with causal realism. However, if, as I aim to argue for here, both mind-independent and mind-dependent norms are crucial for building the correct sorts of models, then interventionism sits in the unrestrictive normativist camp, making it incompatible with causal realism.

Determining whether interventionism is a realist causal project does not merely settle a dispute within philosophy. A third motive for engaging in this discussion derives from the implications of applying interventionism outwith the philosophical domain. Take its application in the scientific domain, for instance. Interventionism and the causal modelling approach is gaining

traction within the natural sciences.[14] Science is typically characterised in terms of its objectivity; indeed, the objectivity of scientific inquiry is often taken to be a good reason for valuing scientific knowledge, and forms the basis for its authority in society. So, depending on what kind of causal project interventionism is engaged in, we might have reservations about its application in such domains. If it turns out that interventionism must rely on mind-dependent norms to deliver causal verdicts, then this would mean that the causal information produced under the view would not be mind-independent/mind-independent*. Basing scientific claims upon such information might therefore partly compromise the scientific ideal.

My aim in this Chapter is to provide reasons in favour of thinking that a successful interventionist theory is one that relies on mind-dependent normative considerations and thus is incompatible with a realist conception of causation. My argument turns on the fact that what makes a causal model a correct model will depend in part upon notions of normality and abnormality, and that sometimes such notions will be determined by and obtain in virtue of minds and perspectives. I'll argue this by taking two widely accepted 'realist' friendly criteria for determining the correctness of a causal model — the stability criterion and the serious possibility criterion — before showing that the criteria are essentially bound up in or need to be supplemented with mind-dependent considerations about what's normal and abnormal. It's worth noting that the idea that interventionism needs to rely on considerations about what's normal has been argued for before by, for example, Halpern (2016), Halpern and Hitchcock (2015) and Hall (2007). But these philosophers do so by arguing that norms ought to compose *additional* constraints on what makes a model correct. Here I'm arguing that extant 'realist' constraints on model construction appeal to normative considerations. In this way, I take my argument in this Chapter to be contributing to the general argument that interventionism must incorporate norms for its success, but for different reasons than have already been given.

Roadmap: In Section 2, I clarify the notion of 'cause' being analysed in this Chapter. In Section 3, I outline Woodward's interventionism. In Sections 4 and 5, I argue that interventionism must supply criteria for what makes a model apt in order to maintain any realist credentials. In Sections 6 and 7, I examine the stability criterion as a principle to determine the aptness of a model. In Sections 8 and 9, I examine the serious possibility criterion as principle to determine the aptness of a model. Throughout these Sections I argue that both criteria draw upon mind-dependent notions of normality. Finally, in Section 10, I outline a response some

[14] Frisch (2014) *Causal Reasoning in Physics*.

interventionists use to pushback against the idea that interventionism is a non-realist project. I then argue that this response is unsuccessful.

3.2 Clarifications

It's becoming increasingly common in the causation literature to distinguish between different kinds of concepts that make up the word 'cause'. Woodward, for instance, differentiates between "type-level", "contributing", "actual", "total" and "direct" causation. In this Chapter, I will focus primarily on the concept of actual causation. Actual causation, or as it is sometimes called "token-level causation", concerns causal statements of the form '*c* caused *e*' where *c* and *e* refer to specific events (or whatever else one takes the causal relata to be). These are the kind of causal statements we typically have in mind when discussing causation. Examples include: 'Suzy throwing her rock at the window caused the window to shatter', 'the rain caused the puddles on the ground' or 'Caesar's crossing the Rubicon caused civil war'. What all these statements have in common is that they refer directly to singular, token events — *Suzy* throwing her rock, *that* window smashing. Actual causal claims are often contrasted with type-level causal claims which take the form '*C* caused *E*', where *C* and *E* refer to event-types. Type-level claims refer to general causal relations rather than specific relations, and they include examples like 'smoking causes cancer' and 'throwing rocks at windows causes windows to shatter'. I focus on actual causation, rather than type-level (or any other conception of the word 'cause'), because most of the authors who have written about normality's effect on causal relations have done so within the realm of actual causation. Hence to talk of causation in some other respect would be to fail to engage with the discussion in a manner that is currently understood. Furthermore, Woodward provides conditions for several causal concepts. As a means of ensuring the scope of this Chapter remains appropriately narrow and directed at the area where analysis of this issue most often takes place, I'll limit myself to talk of actual causation, and largely set aside Woodward's other causal concepts (expect when such concepts are needed to determine actual causation).

Furthermore, following the locution in the previous Chapter, I will refer to defaults and deviations as synonymous with normality and abnormality; a default is a normal state of affairs, and a deviation is a departure from the normal state of affairs.

3.3 Interventionism

The roots of interventionism can be traced as far back as Collingwood (1940),[15] but it began to gain serious traction following the computer scientist Judea Pearl's book *Causality* (2000). In the book, Pearl defends the philosophical claim that causation has to do with what would happen under ideal, surgical experimental manipulations or interventions. But he moves far beyond philosophical generalities to provide a systematic, unified methodology for modelling what would happen under manipulations. Pearl forcefully makes the case for using causal Bayes nets — essentially directed graphs and structural equations — as the formal apparatus from which we can make causal inferences. It is Pearl's formalisation of causation through causal Bayes nets that has had a longstanding impact on the way we think about causation.

Still, it was James Woodward's book *Making Things Happen: A Theory of Causal Explanation* (2003) which popularised interventionism in the philosophical domain. Woodward's book helps makes the forbiddingly technical literature of Pearl's causal Bayes nets accessible to philosophers, but importantly, he also deconstructs the theory, subjects the underlying causal concepts to careful philosophical scrutiny, and rebuilds the view to produce an extremely rich and cogent account of causation and causal explanation.

Like other interventionists, Woodward takes as his point of departure the idea that causal relationships are relationships that are potentially exploitable for the purposes of manipulation and control. Roughly, if *X* is a cause of *Y*, then I should be able to manipulate *X* in the right way that would bring about a change in *Y*. In this way, causal relationships are thought to be relationships of *dependency* potentially exploitable for manipulation and control — *X's* causal status in regards to *Y* depends upon how *Y* reacts under changes to *X*. Woodward's account takes the dependency relation to be one that holds between variables and their values. Variables can be taken to represent one's preferred choice of causal relata — events, facts, properties, instantiations etc. Whether one variable is a cause of another is determined by whether some manipulation on the first variable changes the second variable; that is, whether a change in one variable makes a difference to another. Manipulations must be of a certain kind, however. They must be "surgical interventions", these are interventions which make "exogenous" and "isolated" changes. A change is truly exogenous if it cuts off part of the structure from other parts that would otherwise determine it. In other words, an intervention on *X* would sever the

[15] According to Woodward (2001).

existing causal relationships between *X* and its ancestor causes. An intervention manifests an isolated change if part of the causal structure is changed without altering other parts of the causal structure (except by way of the very intervention). Importantly, these surgical interventions should not be understood exclusively as actual interventions, but rather counterfactual ones. When testing to see if relationships are exploitable for manipulation, we're asking if an intervention *were* to take place, what *would* happen *if* we intervened. Causal relationships between variables thus carry a counterfactual commitment: they describe what the response of *Y* would be if a certain sort of change in the value of *X* were to occur.

The notion of surgical intervention is central to Woodward's view and could thus do with further elaboration. To illustrate the idea further, consider Alex Prescott-Couches helpful description:

> The easiest was to understand the concept of a surgical intervention is to begin by contemplating God. Say God wants to get something done in the world—for example, He wants to turn on the overhead light in my currently unilluminated room to wake me up. How could He go about making this happen? Being God, one thing he could do is swoop in and *directly intervene* on the overhead light without disturbing other aspects of the situation. Such an exogenous and isolated change is an example of what manipulationists [interventionists] call a "surgical intervention" (or, equivalently, simply an "intervention"). An intervention is like an ideal manipulation of a part of the world, the kind of manipulation scientists try to perform when designing controlled experiments. (2017, p. 488)

Woodward's account is seen as a contribution to the Lewisian project of trying to analyse causation in terms of difference making and counterfactual dependence. In his 1973 paper 'Causation', Lewis states that "[w]e think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it" (p. 557). Lewis captures this idea through a counterfactual analysis of causation where *X* is a cause of *Y* if it's the case that had *X* not occurred *Y* would not have occurred. Woodward's interventionism also takes causation to be a difference-making relation constituted by counterfactual dependence, but unlike Lewis, Woodward's counterfactual test looks for changes in the values of variables following combinations of interventions, as opposed to merely testing for the non-occurrence of the effect event following the non-occurrence of the cause event.

3.3.1 Causal Modelling: Structural Equations and Directed Graphs

Following Pearl, Woodward pairs his interventionism with a formal framework of causal Bayes nets. These comprise of systems of structured equations and directed graph, which taken together, create causal models representative of causal relationships. Directed graphs consist of an ordered pair ***{V, E},*** where ***V*** is a set of variables representing the causal relata**,** and ***E*** is a set of directed edges (arrows) representing the causal structure by way of connecting the causal relata. Structural equations, on the other hand, define the causal structure between the variables in the model. Since the encoded relations are causal, the structural equations are asymmetrical. This is to say that the values of the variables on the left-hand side of the = symbol are determined by what appears on the right-hand side but not vice versa.

To illustrate the approach, consider how we can use structural equations and directed graphs to depict this simple causal story: Cat pounces at the fly, Cat traps the fly, and Cat eats the fly. Our story has three salient parts which we can be represented by three variables: CP, CTF and FE. CP = 1 corresponds to Cat pounces, CP = 0 corresponds to Cat does not pounce; CTF = 1 corresponds to Cat traps the fly, CTF = 0 corresponds to Cat does not trap the fly; FE = 1 corresponds to the fly's being eaten, FE = 0 corresponds to the fly's not being eaten. The causal structure is thus defined by the following equations: CP = 1, CTP = CP, FE = CTF.[16] Combining these equations with directed edges indicating the rough causal relationship between variables, we get the following model:

[16] In this thesis I will follow Hitchcock's (2007) notation for constructing the structural equations which makes use of first order logical connectives.

Structural Equations:	Actual Values:
CP = 1	CP = 1
CTF = CP	CTF = 1
FE = CTF	FE = 1

(Figure 1 in Chapter 3)

With the model set-up we can now test to for causation. Let's check whether the Cat's pounce is a cause of the fly's being eaten. Recall that for the interventionist, whether something is a cause depends on whether an intervention on it would bring about an associated change in the effect, so to find out if Cat's pounce was a cause of the fly's being eaten we need to 'wiggle' the variable CP to see if any changes are brought about in FE. From the model we can see that a surgical intervention on CP, say turning it from its actual value of 1 to 0, will bring about an associated change in FE. Namely it will turn FE's value from its actual value of 1 to 0 (since the structural equation tells us FE = CP). As a result, there's a causal relationship between Cat's pouncing and the fly's being eaten (the precise nature of the causal relationships represented in such graphs is outlined a little further on). The causal structure in this story is fairly easy to grasp without these equations and graphs, indeed the causal model appears to needlessly complicate matters. But the power of interventionism and causal models becomes apparent when we consider more complicated structures such as troublesome cases of redundant causation.

3.3.2 Redundant Causation

Redundant causality occurs whenever we have an abundance of causes for exactly the same effect and each of the causes is sufficient for the effect, but none are individually necessary. The kind of causality comes in two variants. First, we have cases of overdetermination, which are symmetric in the sense that all candidate events have an equal claim to be considered the cause of the effect. For example, suppose Billy and Suzy each throw a rock at the window, and each rock reaches the window at exactly the same time causing it to break. Imagine that neither Billy's throw nor Suzy's throw was necessary to break the window — one throw would have

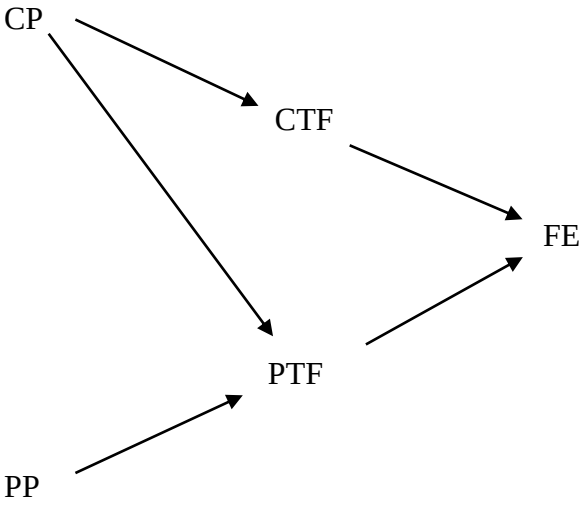
been sufficient. While there's controversy over whether both Billy and Suzy are causes of the smashed bottle, it would seem wrong to single out only one of them over the other as the cause of the effect. The second type of redundant causality are cases of pre-emption. These are asymmetric in the sense that the candidate causes are not on par. Imagine that Suzy's rock reaches the window a couple of seconds before Billy's so Billy's rock ends up passing through an empty space were the glass used to be. As with the first scenario, neither Suzy nor Billy's actions are necessary for the window breaking. Had Suzy not thrown when she did, Billy's rock would have smashed the window anyway. But unlike the first scenario, we do not think that both agents are equally causally efficacious, Suzy is obviously the cause of the smashed window.

Redundant causality poses a well-known challenge to counterfactual theories of causation which identify causes as things which make a difference to the occurrence of their effects. Simply put, the problem is that in cases of redundant causality, the outcome appears to be counterfactually *independent* of any particular cause due to the presence of the would-be cause. Woodward's interventionism is said to possess a significant virtue over other counterfactual accounts because unlike its cousins, the interventionist approach can handle redundant causality cases. To illustrate, consider a more complicated causal story involving Cat. As before let's suppose that Cat pounces at fly, Cat traps fly and the fly gets eaten. This time let's add a second character, Possum, who also pounces to catch the fly, but Cat, being more agile and faster than Possum, gets to the fly first and eats it. But had Cat not got to the fly first, Possum would have trapped and eaten the fly in exactly the same way.[17] This is an example of late pre-emption. Late pre-emption cases are widely considered to undermine the simple counterfactual account, a point which has been recognised by Lewis (2000) himself who consequently abandoned it in favour of his theory 'causation as influence'. The problem is that there is no counterfactual dependence between the effect and the cause when there is a pre-empted backup to the actual cause. As a result, the simple counterfactual account confers the verdict that the actual cause is not a cause at all. In the case just sketched, Cat is not a cause of the fly's being eaten because, given the presence of Possum, it's not true that had Cat not pounced, the fly would not have died. Given the significance of late pre-emption cases in the literature, and the damage it has dealt to the simple counterfactual approach, it is useful to see how the interventionist approach fares in handling such cases.

[17] This example is based on one by Paul, L, A (2000, p.207)

To represent the new causal structure, we need to introduce a couple more variables into the model. Let PP = 1 when Possum pounces, let PP = 0 when Possum does not pounce, let PTF = 1 when the fly is trapped by Possum, and let PFT = 0 when the fly is not trapped by Possum. The situation is represented in the following model:[18]

[18] There is some controversy surrounding the best way to model late pre-emption cases. The debate centres around whether there should be a causal connection represented between the two causal paths {CP-CTF-FE} and {PP-PTF-FE}, and if there ought to be, where the causal connection should be placed. In the model I've drawn here, I have represented a causal connection between the two paths by drawing an arrow from CP to PTF. This is motivated by the thought that when CP is set to 1, it brings about an associated change in the value of endogenous variable PTF, namely it switches PTF off. In other words, Cat pouncing is a direct cause of Possum not catching the fly.
Alternatively, the two causal paths could be connected at a different stage of the causal process by placing an arrow joining the variables CTF and PTF variables. This indicates that CTF (not CP) is a direct cause of PFT. But some have chosen not to represent the causal connection between the two paths at all. For instance, Hitchcock (2007) decides not to draw an arrow between any of the variables along the two causal paths (pp. 526-527). This turns the late pre-emption causal structure into a symmetric overdetermination causal structure.
I have chosen to connect the two causal paths in the way I have because Woodward (2003) does so when modelling a similar case of late pre-emption (p. 79). For the purposes of this Chapter, not much hangs on where we draw the causal connection between the two paths, though I do think such a connection should be represented. Following Woodward's definition of a direct cause (which I describe below), omitting any kind of causal link between the two paths in the way Hitchcock does would fail to accurately reflect the situation's causal structure.



Structural Equations:	Actual Values:
CP = 1	CP = 1
PP = 1	PP = 1
CTF = CP	CTF = 1
PTF = PP ∧ ¬ CTF	PTF = 0
FE = CTF ∨ PFT	FE = 1

(Figure 2 in Chapter 3)

Now to see how Woodward's interventionism handles late pre-emption cases (and other redundant causality scenarios), I need to introduce some of Woodward's causal concepts. I begin with his notion of actual causation which is as follows:

> **(AC1)** The actual value of X = x and the actual value of Y = y (where X and Y are variables in the model)
>
> **(AC2)** For each directed path *P* from *X* to *Y*, fix by interventions all direct causes *Zi* of *Y* that do not lie along *P* at some combination of values within their redundancy range. Then determine whether, for each path from *X* to *Y* and for each possible combination of values for the direct causes *Zi* that are not on this route and that are in the redundancy range of *Zi*, whether there is an intervention on *X*

> that will change the value of *Y*. (AC2) is satisfied if the answer to this question is "yes" for at least one route and possible combinations of values within the redundancy range of *Zi*. (2003 p. 84)

Let's do some unpacking of the definition. **AC1** simply states that the value of *X* and *Y* are assigned the value given by the scenario, that is, their actual value. In regards to **AC2**, three concepts need special attention: direct cause, directed path and redundancy range. Firstly, a direct cause:

> **(DC)** For *X* to be a *direct cause* of *Y* with respect to some variable set ***V*** is that there be a possible intervention on *X* that will change *Y* (or the probability distribution of *Y*) when all other variables in ***V*** besides *X* and *Y* are held fixed at some value by interventions. (2003, p. 55)

Here, ***V***, as before, refers to the variables within the model. *X* is said to be a direct cause of *Y* when we fix all other variables in ***V*** at some value, and there is still a change in *Y* when we manipulate *X*. **AC2** also talks of a *directed path P* from *X* to *Y*. A directed path is a sequence of direct causes running from *X* to *Y* — every variable is a direct cause of the descendent variable. Finally, Woodward states that the other direct causes *(Zi)* that do not lie along the directed path *P* should be fixed according to their redundancy range. Given some actual causal situation, variables that are not on the directed path may be set to non-actual values on the condition that this does not change the value of the putative effect, the range of values that meet this condition is called the redundancy range.

With the definition in place, we can see what interventionism says about late pre-emption. Let's look first at the claim that Cat's pounce is a cause of the fly's being eaten. To do so first focus on the directed path {CP-CTF-FE} and assign CP and FE their actual values as directed by **AC1.** Next fix all other direct causes which do not lie on this path at some combination of values within their redundancy range. The only other direct cause on this model is PTF, which has two values within its redundancy range — 0 or 1 — neither of these values will change the value of the actual effect FE. With the values identified, we can determine whether CP is a cause of FE by considering if an intervention on CP will bring about a change in FE given the possible combination of values PTF might take. First let PFT take its nonactual value of 1, then intervene on CP changing its value from 1 to 0 (Cat does not pounce). This would not bring about an associated change in FE given that PTF is set to 1, hence CP is not yet identified as a

cause. Now let PFT take its actual value of 0, then once again intervene on CP changing its value from 1 to 0 (Cat does not pounce). This would bring about an associate change in FE. **AC2** requires that *X* changes *Y* under *some combinations* of values within the redundancy range of *Zi* along at least one path. In other words, so long as an intervention on CP causes a change in FE under one value of PTF within its redundancy range, then CP is an actual cause of FE. Hence, according to Woodward's definition of actual causation, Cat's pouncing is a cause of the fly being eaten. So far so good.

It's also worth pointing out that there is a second direct path from CP to FE: {CP-CTF-PTF-FE}. According to this path CP would not be a cause of FE. To see this first let CP and FE take their actual values of 1 (**AC1** is met). There are no other direct causes of FE that do not lie along this path (PP being a direct cause of PTF but not of FE), so there's no need to fix any other variables in the model. Next to determine if CP is a cause, we should intervene changing its value from 1 to 0. When CP takes the value 0 it does not bring about an associated change in FE, because when PP is set to 1 the other directed path {PP-PTF-FE} would run to completion. Possum would pounce, trap the fly, then eat the fly, meaning FE would continue to take its actual value of 1. But this outcome does not mean CP is not a cause overall. CP is a cause according to the other directed path we looked at {CP-CTF-FE}, and since Woodward notes that **AC2** is satisfied so long as *X* is a cause of *Y* along at least *one route*, CP is a cause overall.

Now to test if Possum's pouncing causes the fly to be eaten. First focus on the other direct path {PP-PTF-FE} and assign PP and FE their actual values of 1. Next identify the redundancy range of any other direct causes of FE which do not lie along this route. In this instance the other direct cause is CTF, which would have a redundancy range of 1 (it cannot take a value of 0 since this would change the putative effect FE). Now we can see that an intervention on PP changing it from 1 to 0 would not bring about a change in FE in light of the fact that CFT is kept at 1. Thus, Possum's pouncing is not an actual cause of the fly being eaten.

It's worth noting that there remain disputes as to whether Woodward's interventionism really does solve the problem of late pre-emption. One pressing worry first articulated by Hall and Paul (2003), and later further developed by Hall (2006), is that although modelling late pre-emption cases like this generally produces the right causal verdicts, they do so by trading on an ambiguity about how to read the underlying counterfactuals. The problem is that it is only in virtue of this ambiguity that the model is able to deliver intuitively correct causal claims. To

briefly illustrate the problem, consider what happens when we're checking whether Cat is a cause of the fly being eaten. Specifically, consider the directed path {CP-CTF-FE}. When checking for causation along this path, we fix the only other direct cause of FE — PFT — at its redundancy values 1 or 0. Suppose we set PFT to 0, we then check for causation by switching off CP to see if this makes a difference as to whether the fly was eaten, which it does. In order to imagine such a scenario, we are supposing that Cat does not pounce, meaning that Cat does not trap the fly, and the fly does not get eaten. Here's the crux of the problem. Whilst imaging this scenario we also have to imagine that Possum does pounce, and that Possum does not trap the fly. So we have to suppose that Possum does not catch the fly despite the fact that Possum pounces and Cat does not trap the fly. It's completely unclear how this counterfactual is supposed to play out. Ostensibly, it is only insofar of the models of late pre-emption encode these sorts of ambiguities about how to read the relevant counterfactuals that the model delivers the correct causal verdict. I'll leave this discussion here for now, and return to it in more detail in Section 10.1. I raise the problem briefly here to make clear that despite interventionism's success the view still has its bugs. With all that said, however, it's crucial to note that this objection is not considered to be a devastating one. Hall (2006) himself notes that this objection poses no deep threat to the interventionist approach, and that there are other strategies available to the interventionist to straighten out the ambiguities embedded in the underlying counterfactuals (p. 623). Even bearing the objection in mind, interventionism is still said to have the upper hand over the simple counterfactual analysis in terms of handling late pre-emption, because the view has the resources to make the types of nuanced distinctions required to discriminate between the actual cause and the pre-empted back-up — something lacking in the simple counterfactual analysis.

3.4 Interventionism, Realism and Model Relativism

With the stage setting out of the way, we can move to examine whether Woodward's interventionism, and in particular his definition of actual causation, is consistent with a realist reading of causation. Throughout his work Woodward is explicit in his commitment to some minimal form of causal realism. In *Making Things Happen* (2003) he states that "a commitment to some version of realism about causation (in the sense that relationships of counterfactual dependency concerning what will happen under interventions are mind-independent) seems to be built into any plausible version of an [interventionist] theory" (p. 118). Elsewhere he reiterates that his view involves no particular metaphysical commitments beyond a very modest realism:

> This modest realism consists in the (I would have thought uncontroversial) assumption that the difference between those relations that are merely correlational and those that are causal has its source "out there" in the world (as philosophers like to say) and is not, say, somehow entirely the result of our "projecting" our beliefs and expectations onto the world with the result that some relationships look causal even though none "really" are. Of course, it is a fact about us and our interests that we value information about relationships relevant to manipulation and control, but it is the world (and not just our interests) that determines which such relationships hold and in what circumstances. (2014, p. 698)

Despite these remarks, it's not immediately obvious that Woodward's modest commitment to causal realism is consistent with interventionism. Woodward's definition of actual causation is relativised to a variable set; *X* might be an actual cause of *Y* with respect to some variable set ***V*** but not with respect to a different variable set ***V*.*** This appears to invite a variable relative interpretation of actual causation, according to which the truth value of causal claims depends upon the variable set against which they are assessed. Hence, whether some variable is an actual cause of another is sensitive to how *we choose* to represent the causal structure. This spells trouble for a realist reading of interventionism. As was demonstrated in the previous chapter, central to the realist position is the thought that whether one thing causes another is independent of our representation of it. Indeed, the account's model relativity has led some to doubt interventionism's realist credentials. For instance, in a review of *Making Things Happen,* Peter Menzies (2006) writes that "the crucial role played by the choice of variables in the representation of causal relations raises some serious questions about the extent to which a full-blooded realism about causation can be sustained. The fact that the correct application of causal concepts depends on certain key representational choices seems to imply that the truth-conditions of causal concepts are not completely objective or mind-independent" (p. 852).

3.5 Apt Causal Models

There is an obvious response to such a worry. One can simply claim there are 'right' and 'wrong' variable sets. The wrong variables sets are those that falsely or inaccurately encode dependency relations that are exploitable for manipulation and control, they therefore produce *false* causal claims. Whereas the correct variable sets encode the right dependency relations potentially exploitable for manipulation and control, and hence produce *true* causal claims. (This does not necessarily entail that there is only one unique model for any given causal story.

For many complex systems a range of different causal models utilising different variables may be appropriate, such that there may be many right models. The point is that there are objectively right and wrong models).[19] In light of the fact that our aim is to discover true causal claims, we ought to construct the right kind of models. This line of reasoning deflects the concern that the truth value of causal claims depends upon how we chose to represent the world; there is still some mind-independent fact of the matter about what causes what, and our job, as modellers, is to produce the model most conducive to uncover those facts.

So two significant implications hang on whether criteria for a model's correctness can be articulated. Firstly, and most straightforwardly, whether or not we uncover *true* causal claims depends upon our choice of model — correct models yield true causal claims, wrong models yield false claims. Secondly, interventionism's realist credentials depend upon whether we can delineate what makes a model correct. For if we cannot establish the objective standards for correctness, then any given model will be as good as any other, and in turn, the causal claims produced under these models will be as good as any other. This invites a model-relative reading of interventionism that does seem to undermine a realist understanding of causation.

The challenge of identifying what makes a model correct has been discussed before, and it is often referred to as the need to establish models which are 'appropriate' or 'apt' for representing causal relationships. Identifying what features make a model appropriate has proven to be a difficult task which has received comparatively little attention. As Paul and Hall (2013) remark: "it is an excellent question, inadequately addressed in the literature, precisely what principles should guide the construction of a causal model" (p. 17).

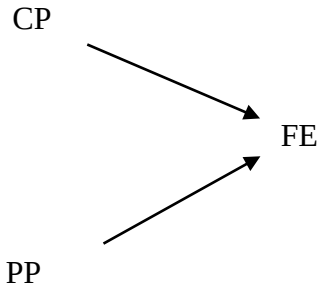
In what follows, I outline two pressing questions about model construction that have presented difficulties in determining a model's aptness; namely, how many variables ought to go into a model? And what values should these variables take? I then explore some principles which have been proposed to answer these questions. I have been careful to select principles which

[19] Some philosophers and computer scientists do not believe that there are correct models for any given situation. Halpern writes that "I am not sure that there are any 'right' models, but some models may be more useful, or better representations of reality, than others" (2016, p. 107). Halpern is happy to give up on realism about actual causation, but for those who want to keep their realist credentials, their causal models must conform to a criterion of correctness that describes how to construct a model in line with causal reality. Indeed, Woodward remarks that "I assume (in agreement with what I take to be the attitude of most scientists) that many choices of variables are, from the point of view of causal analysis, defective in some way—among other possibilities, they may lead to the formulation of causal claims that are false or impede the discovery and formulation of true claims." (2016, p. 1051).

are supposed to be realist friendly in the sense that their application is thought to be objective, impartial and independent of any capricious anthropocentric judgements. The principles are endorsed by interventionists who claim that interventionism is compatible with realism (such as Woodward 2003, 2014), as well as those who are not interventionists but nonetheless agree that interventionism is compatible with realism (such as Blanchard and Schaffer 2017). However, I aim to show that once these purported realist friendly conditions are spelled out in sufficient detail, they actually bottom out into requiring the use of normative considerations. Specifically, they require an appeal to mind-dependent considerations about what's normal and abnormal. I'll thus argue that *apt* causal models are sensitive to mind-dependent normative considerations, and if this is true, then an interventionist framework which employs causal models will be incompatible with causal realism. In terms of the taxonomy outlined in the previous chapter, my argument would entail that interventionism is a view which exemplifies unrestricted normativism.

3.6 How Many Variables Should be Expressed in a Model?

The first aspect of model construction that continues to present difficulties for the modeller concerns how many variables ought to go into a model. It's commonly thought that one should have enough variables in a model to sufficiently represent the essential causal structure of the situation, but striking a balance between sufficiently expressing the causal structure and not over saturating the model with information is a tricky task. Initially one might be inclined to think that the actual situation should be expressed through the fewest variables as possible given that generally it is preferable not to unnecessarily overcomplicate things. Furthermore, it is plausible to suppose that fewer variables would likely glean clearer more succinct causal information. However, too few variables can under-express the essential structure of the actual situation which in turn can generate intuitively incorrect causal claims. To illustrate consider the late pre-emption Cat and Possum case discussed earlier. Recall that Cat and Possum both pounce to catch a fly, Cat being more agile and faster than Possum gets to the fly first and eats it. But had Cat not got to the fly first, Possum would have eaten the fly in exactly the same way. The model I used in Figure 2, demonstrated that Woodward's definition of actual causation was able to adequately identify Cat (and not Possum) as a cause of the fly's death. However, had I used a model with fewer variables the verdict would have been different. Consider this alternative representation of the scenario:



Structural Equations:	Actual Values:
CP = 1	CP = 1
PP = 1	PP = 1
FE = CP ∨ PP	FE = 1

(Figure 3 in Chapter 3)

In Figure 3 I have omitted to include two variables: CTP representing that the fly is trapped by Cat and PTF representing that the fly is trapped by Possum. In Figure 3 Cat remains an actual cause of the fly's being eaten since an intervention on CP makes a difference to FE when PP is set at one value of its redundancy range, namely 0. However, according to Woodward's interventionism, this model wrongly recognises Possum as an actual cause of the fly being eaten. To see why first focus on the directed path {PP-FE} and assign PP and FE their actual values of 1. Next identify the redundancy range of any other direct causes not on this directed path, which in this case is CP. Unlike Figure 2, CP's redundancy range is both 1 and 0 because CP's taking either of these values would not bring about a change in the putative effect. When CP takes the value 0 an intervention on PP will bring about a change to FE, and thus we get the verdict that Possum's pouncing is a cause of the fly being eaten. The verdict is not surprising; there's nothing in Figure 3 that differentiates CP from PP; hence they end up playing symmetric causal roles.

3.7 The Stability Criterion

When presented with these two representations of the causal story, presumably a defender of interventionism is going to stress that we should prefer Figure 2's expression over Figure 3, but what criteria can the interventionist appeal to in order to justify this preference? Here's one suggestion. In his 2016 paper, 'The Problem of Variable Choice', Woodward argues that we ought to build models which contain stable causal relationships (p. 1053). The stability of a

causal relationship concerns the number of background conditions that must be preserved for it to occur, where the relationships with fewer such background conditions are thereby more stable. Hence, a stable causal relationship is one that holds across many varied background conditions, whilst an unstable relationship is that one that would hold in only very few background conditions. To illustrate, suppose that a computer is programmed to turn on when a certain button on the computer's keyboard is pressed. The causal relationship between pressing the button and switching on the computer is relatively stable, for the causal relationship will hold in many and diverse background circumstances. It will only fail to hold in circumstances where certain internal elements of the computer's system are defective. Testing for stability in a model requires that we hold fixed the variables that encode a causal relationship whilst manipulating various background conditions to see if this relationship continues to hold in these various background conditions.

Stability is relevant to decisions about the choice of variables because some number of variables may enable more stable causal relationships than others. This is to say that how we choose to represent the causal relationships in the model determines whether such relationships hold across many or few background conditions. Hence, Woodward argues that we should "look for variables that allow for the formulation of causal relationships that are stable in the sense that they continue to hold under changes in background conditions" (p. 1055). (I take a closer look at the notion of causal stability in Chapter 5, but this coarse characterisation will do for the purposes of this discussion).

The reason for choosing variables which produce stable relationships has to do with the interventionist's core claim that causal relationships are relationships potentially exploitable for manipulation and control. Stable connections can be exportable from one set of circumstances to another, they are therefore more generalisable, and thus provide more information relevant to manipulation and control.[20]

[20] Several other philosophers and computer scientists such as Halpern and Hitchcock (2010) also emphasise that there is a requirement of stability on apt models. But how these authors delineate the stability constraint is slightly different from Woodward. For example, Halpern and Hitchcock argue that stability has to do with whether the model can preserve its original causal verdict. They argue that if the addition of variables into a model produces causal verdicts that are different from that of the original model, then the model is unstable. Whereas, if the addition of variables in a model produces the same causal verdict as the original, then it is stable. They argue that it preferable to have stable models over unstable ones (p. 395). I set aside this understanding of stability and focus on Woodward's definition chiefly because Woodward's interventionism is the focus of this Chapter, and

Following the stability constraint lends support to preferring Figure 2 over Figure 3. In Figure 3 the causal connection between CP (Cat's Pounce) and FE (the fly's being eaten) is relatively unstable insofar as any number of changes in the background circumstances could lead to a breakdown in the connection. For example, alterations to the speed at which the animals pounce, the time at which they pounce, their dexterity or prowess could all result in a severing of the causal relationship between Cat's ponce and the fly being eaten. By contrast, the addition of the two variables CTF and PTF in Figure 2 creates at least one causal relationship which is extremely stable, namely the relationship between CTF (Cat trapping fly) and FE (the fly being eaten). Unlike, the causal connections in Figure 2, the link between the fly's being eaten by Cat and Cat trapping the fly will continue to hold across a range of changes to the background conditions. Alterations to Cat and Possum's speed, agility, timing and so on will not sever the connection. As a result, requiring that apt models express stable causal relationships promises to provide a justification for the interventionist to prefer Figure 2 over Figure 3.

However, it's not clear that a *causal realist* would be entitled to appeal to notions of stability in order to justify the aptness of a model. This is because a closer look at stability reveals that it requires an appeal to the full range of normative considerations in order to measure it. Testing for stability entails holding fixed the variables in the model whilst altering the background conditions, this requires two specific steps: (a) one must differentiate between variables and background conditions, and (b) one must select the set of background conditions relevant for manipulation. Completing the second step — selecting the background conditions suitable for manipulation — is crucial for measuring the kind of stability that makes a model apt. i.e., stability that provides information relevant for manipulation and control. Settling the first step — differentiating between background conditions and variables — is not only crucial for testing for stability, it is also crucial to the interventionist project more generally. A modeller is required to make a choice which distinguishes between the variables she puts into the model from the background circumstances omitted from the model. A failure to correctly distinguish between these two entities can lead to models which produce intuitively incorrect causal verdicts. I will show that successful strategies for settling both of these steps appeal to factors pertaining to what's normal and abnormal, and further I'll demonstrate that sometimes considerations about what's normal and abnormal relevant to measuring stability depend upon

because I come back to look at Woodward's notion of stability in Chapter 5, focusing on it now is thus a useful primer for the discussion to come.

minds. Thus, we have reason to doubt that the causal realist can appeal to stability to justify the aptness of a model.

3.7.1 Differentiating between Variables and Background Conditions

I won't dwell on this step since I have discussed it in the previous chapter. There I noted that we regularly draw a distinction between the cause of an effect and the background conditions which made the effect possible. If a forest fire occurs, the forest ranger will select the lightning strike as a cause of the fire and relegate the presence of oxygen to a background condition. If we want to reflect these kinds of judgments in our causal models, then we must find a principle that distinguishes causes from background conditions. For notice that if we do not find such principle then what we think of as background conditions will likely come out as bone fide cause under the interventionist framework. For instance, with regards to a variable set that includes lightning strikes, forest fires and oxygen, oxygen would likely come out as a cause of the fire since manipulating its presence would bring about an associated change in the occurrence of the fire. The particular challenge for the modeller then, is to represent potential causal relationships in the model through variables, whilst screening-off any background conditions from entering into the model. This is required to test for stability.

To meet this challenge, it is becoming orthodoxy to invoke the distinction between normal and abnormal states. Normal states constitute background conditions and are therefore omitted from the model, whilst abnormal states are the types of things that make-up the essential causal structure and hence are represented as variables in the model. An apt model would not include oxygen since its presence is a normal state. I already addressed why invoking notions of normality and abnormality can mark the intuitive difference between causes and background conditions in Chapter 2. With that said, I will note a potential virtue of deploying this strategy which is specific to the modelling methodology.

Some philosophers understand the concept of causation to be context-sensitive.[21] Roughly causation as a context-sensitive concept entails that the truth conditions for causal claims are partly determined by the context of the situation, meaning that essential causal structure does not solely determine the truth value of causal claims. In this way, we could have the same causal structure over two different contexts and end up with two different sets of causal claims. To illustrate, consider Hilary Putnam's (1982) well-known vignette where he describes a pair

[21] Peter Menzies (2004) is probably the most well-known advocate of this view.

of alien Venusians who land on Earth, observe a forest fire and say “I know what caused that — the atmosphere of the planet is saturated with oxygen” (p. 150). Despite the fact that the objective causal structure of the situation is shared by us humans and the Venusians, the Venusians’ contextual parameters lead them to promote oxygen from a mere background condition to an actual cause. If a modeller wanted to reflect the idea that causation is context-sensitive, then appealing to the difference between normality and abnormality provides a promising avenue. For the Venusians, the presence of oxygen is a deviant state of affairs, and hence, an apt model will express it as a variable. Switching the oxygen variable ‘off’ will likely bring about an associated change in the occurrence of the fire, therefore the model would reflect the fact that, from the Venusian modeller’s perspective, oxygen is an actual cause.

The forest fire case is a canonical example used to highlight the difference between background conditions and actual causes, however it does not provide support for my argument that the stability criterion is a non-realist friendly criterion. This is because the norms being invoked to distinguish between actual causes and background conditions in the case are mind-independent. The presence of oxygen represents a normal state of affairs on account of the fact that its presence is statistically normal, there's no mind-dependent sense as far as I can see which makes the presence of oxygen a normal occurrence in earth’s atmosphere. Consequently, to show that the stability criterion is incompatible with a realist approach, we need to search further afield for a case in which mind-dependent norms are making the relevant distinctions. As it happens, I provided such an example in the previous chapter. Suppose Hollie is trying to park her car in a busy supermarket carpark. After doing laps of the carpark, she finds a spot, and parks her car. But in her rush Hollie forgets to buy a parking permit, and on her return, she notices a parking ticket on her windshield. I said that it was plausible to suppose that Hollie’s failure to buy a parking permit caused her to receive a parking ticket. Or perhaps we might also want to say that Hollie’s parking the car where and when she did caused the parking ticket. I also said that it would presumably be incorrect to claim that Hollie’s ticket was caused by all the other vehicles in the carpark being parked within the white lines. For surely this is a mere background condition to Hollie being issued a parking ticket. Further, I argued that appealing to a normative framework allows us to make this distinction because it is normal for motorists to park within the white lines, hence we’re permitted to relegate this aspect of the story to a background condition. But notice that the specific normative consideration being invoked here is a social norm, and hence, a mind-dependent norm; parking within the white lines is socially

normal behaviour.[22] As a result, if one wanted to build a model in accordance with the stability criterion, one must invoke normative considerations which would undermine a realist conception of causation.

This is all to say that distinguishing between the causal relations and the background conditions which made those causal relations possible requires an appeal to the difference between what's normal and what's abnormal. And further, sometimes conceptions of what's normal and abnormal will comprise of mind-dependent norms. This in turn means that in order to measure stability one must appeal to features that are incompatible with a realist project.

3.7.2 Relevant Background Manipulations

Even if there was a realist friendly way to distinguish between background conditions and variables, there is another point at which normative considerations enter the picture when measuring stability. Once the difference between background conditions and variables has been established, measuring stability requires that we alter the background conditions to test whether the actual causal relationship encoded by these variables would continue to hold across different circumstances. What kind of alternations should we make? A lot of potential interventions to the background conditions will be uninteresting from the perspective of testing for stability. Let us return to the forest fire case. Suppose we're looking to see how stable the causal relationship is between the lightning strike and the forest fire, and suppose that just as the lightning hits the tree, someone sneezes 500 miles away. Testing to see whether the causal connection between the fire and the lightning strike would hold given an intervention on whether the person sneezes is extremely uninteresting, presumably because it is easy to foresee that, if the world remained the same in all other respects, an intervention on the sneeze would not lead to changes in the causal relationship between the lightning and the fire. Other interventions are equally uninteresting but for different reasons. We could alter the background circumstances by coating the tree in a magical potion which prevents wood from catching fire. Although this time the alteration would make a difference to the causal relationship (the lightning strike would no longer cause a forest fire) we nonetheless regard this information as irrelevant.

[22] There might be a sense in which parking within the white lines is statistically normal, but the statistical norm seems to be derived from the fact that parking within the white lines is first and foremost adherence to a prescriptive rule; one *ought* to park within the white lines.

Other alterations, however, strike us as more interesting. For instance, it seems sensible to consider whether the forest fire would have occurred had there not been unusually strong winds that day, or had the weather not been persistently hot and dry in the months leading up to the lightning. What principle makes *these* background alterations relevant but not others? There has been very little discussion about the kinds of circumstances we ought to intervene on to test for stability. Here I'll suggest one principle for determining such relevance. The reason why some interventions strike us as relevant has to do with the reasons for which stability was introduced as a criterion in the first place. Recall that the stability constraint is motivated by the thought that stable relations provide useful information relevant to manipulation and control. Stable connections can be exported from one set of circumstances to the next, hence they're more generalisable, and therefore offer information about how one can predictably manipulate that kind of causal connection. Given this motivation, I suggest that the changes which are relevant are those that reveal information about how we can control the causal relationships in the model. We aren't concerned with whether the forest fire would occur had the tree been covered in a magical resin, because given our lack of access to magic resins a change of this type does not offer useful information about how to control the causal relationship. The same goes for other alterations that represent abnormal, confounding or extraordinary changes, because they too reveal little information about the nature of actual causal relationships and how we can control them. For example, it would be of little use to us from the perspective of manipulation and control to consider what would have happened to the fire if a large group of rangers carrying ladders and buckets of water happened to be passing by the tree at the exact time the lightning struck. By contrast, checking for stability by making changes to the circumstances which make them fairly normal, ordinary or expected does seem to offer information about how one could manipulate the world — it tells us what kind of interventions can change a causal connection in the circumstances we typically find ourselves in. This is all to say that since we're interested in how to manipulate causal connections it makes sense to consider those causal connections in a context *in which we normally occupy*. Thus, the relevant kinds of changes to the background conditions are those that either keep the background conditions normal (if indeed the actual circumstances are normal) or change the abnormal background conditions to normal conditions.

I recognise that sometimes a 'normal manipulation' in the background circumstances doesn't obviously yield information we can exploit for manipulation and control. A move to consider what would happen to the causal connection between the fire and the lightning if the weather

had been less dry and hot than it actually was, would be, let's suppose, a move to consider what would happen under more normal background conditions. But supposing we found that the causal relationship between the fire and the lightning were to break following an intervention on the weather, one might doubt that this is useful information. After all, we can't manipulate the weather. Whilst it's true that making *direct* interventions on the weather is beyond our capacity, finding out that the causal relationship would break if we were to make such an intervention yields useful information about how we can *indirectly* manipulate the world. Suppose we already know that global warming increases the risk of hot, dry weather, and given that the model has told us hot, dry weather was an actual cause of the forest fire, we can come to the conclusion that global warming causally contributes to forest fires. Now this information is useful from the perspective of manipulation; changing behaviour to mitigate global warming will reduce the kinds of weather states conducive to causing forest fires.

Importantly, the sense of normality we are dealing with when it comes to selecting which backgrounds conditions are relevant for manipulation comprises of the full range of normative considerations including those that are mind-dependent. That is, when measuring stability, we are seeking to make 'normal manipulations' on background conditions, where what's normal encompasses mind-dependent norms. To illustrate, consider Hollie again. Suppose there is a causal relationship between Hollie forgetting to purchase a parking permit and Hollie being issued a ticket. To measure how stable this causal connection is we need to consider whether the connection would hold under various departures from the actual background conditions. Relevant departures to consider might include how many parking attendants were working at the time, since had there been fewer, Hollie might have returned to her car in time to avoid a ticket. This is useful information in terms of control if Hollie wants to avoid getting a ticket. Irrelevant departures, on the other hand, will include changes to the fact that all the other vehicles in the carpark are parked within the white lines. It might be true that Hollie would not have been able to park, and hence would not have received a ticket, had all the other vehicles been inefficiently scattered around the carpark. Still envisaging such an alteration strikes us as irrelevant when it comes to determining stability. This is because we are interested in finding interventions that would allow us to control the causal story in circumstances that we normally find ourselves in. Given this orientation, asking what would have happened had people parked abnormally provides relatively useless information in terms of being able control the world around us. And notice that the kind of normality being invoked here to identify what counts as

a normal manipulation is one derived from a social norm; people parking their cars within the white lines is socially normal behaviour.

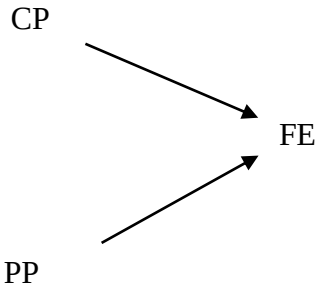
I'll end my discussion of the stability criterion here. Hopefully the discussion has served to illustrate that despite the fact stability is often invoked as a constraint on how many variables one ought to put in a model, there has been very little work done to develop what measuring stability would actually entail. Firstly, it is unclear how one distinguishes between background conditions and variables, and even if this issue is settled, it is unclear which background conditions are appropriate for manipulation in order to test for stability. I've tried to show that invoking a normative framework which makes use of the distinction between normality and abnormality can solve these issues. And further, that the relevant sense of normality and abnormality being appealed to includes mind-dependent norms. If one wants to invoke stability in order to determine a model's aptness then, the resultant view would be in line with unrestricted normativism.

3.8 What Values should the Variables take in a Model?

The second difficulty I'll focus on regards the values that the variables in the model ought to take. A modeller has considerable leeway in allocating the values to variables, but much like with regards to our choice of variables, one must be careful to choose values which accurately represent the causal story. For simplicity I have thus far considered variables that take the value 0 or 1 depending on whether the thing represented by the variable occurs or not. Nonetheless, to fully express the situation, a variable might need to go beyond binary values. Specifically, it might need to take on different sorts of states other than 'on' or 'off', and it might need to take on more than two states. For instance, in 'Iona's saying hello loudly caused Jim to jump', we could let *H*, representing Iona's saying hello, take the value of 1 or 0 depending on whether she says hello or does not say hello. Turning *H* from 1 to 0 would make a difference as to whether Jim jumps, therefore making Iona's saying hello a cause of Jim's jump. But a model like this — which only makes room for the possibility of saying hello or not saying hello — fails to capture the thought that it is her saying hello *loudly* that causes Jim to jump.[23] An apt model will be one where there's a possibility to intervene on the loudness of Iona's greeting.[24]

[23] It remains a live debate in metaphysics as to whether saying hello and saying hello loudly are different events, but debates about event individuation are not especially relevant to interventionism's causal modelling since variables and their values don't necessarily map onto the ontology of events.
[24] It's also worth pointing out that the fact that interventionism allows a variable to take more than the binary values of on or off is seen as a significant virtue of the view. It can, in theory, make the type of

This discussion might suggest a relatively fine-grained solution to the problem of value allocation. Rather than allowing interventions to merely change the status of variables from 'occurs' to 'does not occur' (and vice versa), perhaps we ought to be more nuanced by accommodating for the fact that interventions can also change the descriptive properties of variables. However, a fine-grained approach isn't always the best way to model the causal picture. To see this let's go back to Cat and Possum. Here's Figure 3 again:



Structural Equations:	Actual Values:
CP = 1	CP = 1
PP = 1	PP = 1
FE = CP ∨ PP	FE = 1

(Figure 3 in Chapter 3)

I argued that this model was not apt because Possum's Pounce comes out as a cause of the fly being eaten (at least by the light of Woodward's definition of actual causation). In the original case, I let FE take the value of 1 or 0 depending on whether the fly gets eaten or not,. Interestingly the causal claims generated by the model are different when we take a more fine-grained approach to value allocation. Let's allow FE to take the value of 0 when the fly does not get eaten or 1 if the fly gets eaten *by Cat*. At first glance these value allocations fare much better than the original Figure 3 allocations at generating correct causal claims. Under the new set of values Cat comes out as a cause of the fly being eaten and Possum does not (this is because CP now has a redundancy range of just 1, since 0 would alter the actual value of FE — fly gets eaten by Cat). Despite the fact that these values get at the intuitively correct causal

nuanced distinctions which allow us to say that it was saying hello loudly and not just saying hello simpliciter that was the cause. This is something that the simple counterfactual analysis of causation cannot do.

claims, there seems to be something dubious about assigning variables values like these. By allowing FE to take the value of 'fly does not get eaten' or 'the fly gets eaten *by Cat'*, we seem to bake into the model the dependency relation that reflects our identification of actual causation. We would thus be building models in bad faith; we already have in mind a causal verdict, and encode into the model values which will preserve this verdict. [25] If a fine-grained approach isn't a successful principle, what other factors can guide value allocations?

3.9 The Serious Possibility Criterion

According to many interventionists, including Woodward (2003, 2016), Halpern and Hitchcock (2010), the values that ought to go into a model are those that represent serious possibilities. Broadly, the idea is that we want variables and the values of those variables to correspond to possibilities we could plausibly see manifesting in the actual context. When my fish dies from lack of food, for example, we don't hold Bono causally responsible for its death, even though it may be true that had Bono visited my flat and fed the fish, the fish would survive, because this possibility is not one we are willing to take seriously, hence it is discounted as a causal factor. Similarly, when investigating whether it was McEnroe's serve that cost him the tennis match, we don't wonder whether the outcome of the game would have been different had he been serving 100mph faster than he would usually, even though it may be true that had he served at this speed, he would probably have won. For McEnroe to serve 100mph faster than he would ordinarily is not a serious possibility, and therefore we don't take his failure to serve at this speed as a causal factor in his loss. In practice, the serious possibility criterion entails that the endogenous and exogenous variables and values in $\boldsymbol{V}$ should only express serious possibilities.[26] Modelling the fish's death would require that we do not include a variable representing Bono feeding the fish, and modelling McEnroe's lost tennis match would entail that the variable representing his serve cannot take a value McEnroe could not reasonably achieve.

The serious possibility criterion is a plausible candidate for building apt models. For one thing, restricting the values to only those serious possibilities prevents the model from generating intuitive correct causal claims through dubious methodology. Had the variable representing McEnroe's serve been allowed to take a value that is 100mph faster than his actual average for

25 When discussing a similar strategy in a different context, Halpern refers to this strategy as 'cheating' (2016, p. 31).

[26] Unless, of course, the actual situation happens to involve non-serious possibilities, in which case the endogenous value in $\boldsymbol{V}$ should take on the actual, non-serious possibilities.

the match, his serve speed would, let's suppose, bring about an associated change in the outcome of the match, meaning that his serve would be an actual cause of his loss. Even if it turns out that this causal claim is true, the methods by which we would have got to it seem to be problematic; we'd be building models in bad faith again. Secondly, the serious possibility criterion directs us as to which omissions ought to be expressed in the model. In the previous chapter, I noted that one pressing challenge for a theory of causation is to differentiate between absences which seem to be causally efficacious and those that seem to be causally inert. For the interventionists, this challenge will manifest by way of needing to find a principle that excludes the causally inert omissions from being represented in the model and includes the salient omissions in the model. The serious possibility criterion offers one way of doing exactly this. Representing Bono's failure to feed the fish is unjustified on the grounds that the event for which is his omission — feeding the fish — is not a serious possibility. Whereas the event which is, say, my own failure to feed the fish, is a serious possibility and its representation in the model would therefore be justified.[27]

The serious possibility criterion has also been defended by non-interventionists. Thomas Blanchard and Jonathan Schaffer (2017) claim that the serious possibility criterion is an "independently justified constraint" (p. 197) that can guide us as what variables and values ought to comprise the variable set. Indeed, the authors claim that the criterion does enough work in this regard so that there's no need to complicate matters by invoking normative considerations to determine the aptness of model (I disagree with this claim, and I'll come back to it later). With regards to the fish case, for example, Blanchard and Schaffer would claim that the causal asymmetry between my omission and Bono's omission is simply a reflection of the fact that Bono's feeding the fish is an absurd possibility. There is no apt model, they argue, in which wiggling whether Bono feeds my fish wiggles the fate of the fish, because there is no apt causal model that considers so ridiculous a scenario as Bono popping by my flat, fish food in hand, to engage in random acts of fish feeding (p. 197).[28] For them, the notion of a serious possibility is all that's needed to diagnose cases where some events or omissions, although counterfactually connected to the effect when exhibited in the model, nevertheless strike us as causally irrelevant.

[27] This is assuming I'm not on holiday, as I was in a similar example used in the previous chapter.
[28] Here I'm paraphrasing Schaffer and Blanchard's reaction to a different case involving absences.

However, I have reservations about the success of the serious possibility criterion. Although helpful, its usefulness is limited. Such heuristics are unlikely to yield a uniquely best choice of values, at most it will rule out certain choices as bad or defective, it therefore has a bearing on what values ought to be *excluded* from the model, but it doesn't help much with regards to what ought to be *included* in the model. To illustrate consider the McEnroe model again. When investigating whether his serve caused him to lose the match, we excluded the variable representing his serve from taking a value that was way above McEnroe's average serving speed in light of the fact that this would be improbable or perhaps impossible. But what values should be included? If we merely followed the serious possibility constraint then the variable representing the serve could take a tremendous number of values; for surely it's a serious possibility that McEnroe could serve anywhere from slightly quicker than the fastest serve he's ever played to the slowest possible serve one can play without incurring a fault. But, of course, including all of these values is not helpful from the point of view of our causal inquiry. Although it might be a serious possibility for McEnroe to serve at 70mph, 71mph, 72mph and so on, testing what would happen to the causal relationships in the model at these values is not very useful or interesting from the perspective of discovering information relevant to manipulation and control of the game. Presumably if he serves at speeds around 70mph his opponent will be able to return his serve, McEnroe will still lose, and we will have failed to garner any information that might be helpful in manipulating the outcome of the game.

To gather information that is useful for the purposes of manipulating and controlling one's environment, we need to restrict the range of variables and values represented in the model, and to do this we need to move beyond the serious possibility criterion. So what other principles could one draw on? Consider the McEnroe model again. To investigate whether his serve cost him the match, first set the variable representing his serve at its actual value, which let's suppose was an average of 108mph. Then consider an intervention on this value that would reveal potentially pertinent information for manipulating and controlling the outcome of the match. Given this end, it seems that an appropriate intervention would be one that changes the value from his actual average serving speed in *that* match to his overall average serving speed, i.e., the speed he would *normally* serve at. If the intervention brings about an associated change in the outcome of the game, then (supposing that the other conditions for Woodward's definition of actual causation are met) McEnroe's serve was an actual cause of him losing the match. If this intervention does not make a difference to the causal relationships in the model, it seems reasonable to then draw a wider net and set the value at whatever would be a normal

serve speed for an elite male player. Let's suppose that this is faster than McEnroe's normal serve speed. Again, if the alteration brings about an associated change in the outcome of the match, then according to interventionism, McEnroe's serve causally contributes to his loss — useful information if McEnroe wants to improve his performance.

The discussion suggests that the kinds of variables and values we take to be relevant for purposes of manipulation and control are heavily guided by considerations of what's normal and abnormal for that context. It's natural to compare the actual causal picture with what would have happened had the pieces of the picture been more normal, since this allows us to pick up on what might be appropriate targets for intervention. To further illustrate, suppose this time we wanted to know whether Martina Navratilova's serve cost her the match. Although the causal inquiry is the same as the one launched for McEnroe's, the kinds of values considered to be relevant are quite different. We wouldn't intervene, for example, by changing her speed to that of McEnroe's or the average elite male tennis player, rather we manipulate the speed to bring it up to her average speed, or the average speed of an elite female player. The point is that the choice of values changes depending upon what we take to be normal for that context in which the causal inquiry is launched.

Once again, the sense of normality I'm invoking here is not one that restricts itself to mind-independent norms but rather draws upon the full range of normative considerations. There will be instances where what we judge to be a normal value will be derived from mind-dependent features of the world. For example, suppose A attends a job interview dressed in a chicken costume, unsurprisingly A is not offered the position. If we wanted to find out what the causes were of A's failing to secure the position, we might build a model which includes variables like A's interview outfit, A's interview technique, and A's interview answers. What kind of values ought we give these variables in order to determine the causal facts? Well, with respect to the variables representing A's outfit, it would be wise to assign it values that correspond to more conventional interview attire. If an intervention turning the variable from its actual value into one representing a more normal value leads to a change in the effect, then we have gained pertinent information about how to influence the outcome of the interview. And if it does not lead to a change in the effect we have still arrived at useful information, for now we know that other variable(s) in the model must have causally contributed to the unsuccessful interview. Critically, what makes the assignment of these non-actual values relevant depends upon a social or cultural norm insomuch as dressing a certain way for

interview is a socially or culturally imposed rule.[29] Also notice that if we relied solely on the serious possibility criterion, we would be left a drift as to what kind of values the variable representing A's outfit ought to take. The criterion would tell us to exclude other non-serious values like dressing in sportswear, but it wouldn't tell us what values we ought to plug-in.

The idea that considerations about what's normal and abnormal ought to influence the values we put into the model has been remarked on before.[30] The specific point I wish to make here is that the serious possibility criterion, although useful insofar as it tells us what values ought to be excluded, needs to be supplemented with a normative framework in order to guide us as to what needs to be included in the model. Moreover, the norms which are useful in guiding us in this respect will include mind-dependent norms.

3.10 An Interventionist Response

Before concluding this Chapter, I'd like to respond to a move that's quite often made in response to the sorts of claims I've just made. Roughly, the response begins by granting that causal facts under interventionism are model-relative, and that our choice of model will depend upon subjective judgements, but to deny that this entails that interventionism is a non-realist theory. Put simply, the idea is to grant that causal claims will be true or false relative to a model and a modeller, but to point out that what makes these claims true or false is determined by patterns of counterfactual dependence which lie behind the model, and since the existence of counterfactual dependence is an objective matter, causal claims ultimately get their truth value from an objective source. To put this response in the terminology I've been using, interventionists can simply accept the argument that I've been making here — namely, that mind-dependent norms determine what we put into a model — whilst maintaining that causal facts are mind-independent, because it is a mind-independent matter whether there is counterfactual dependence between the values of variables in the model. Insofar as this is true, interventionists claim that their theory can retain its realist credentials.

Woodward takes up this line of defence. According to his definition of actual causation, of what it is for X to be a cause of Y is relativised to a set of additional variables $\boldsymbol{V}$. X may be a

[29] Again, there's a case to be made that dressing in a chicken costume for an interview violates a statistical norm as well as a social norm, insomuch as dressing in a chicken costume is statistically unusual. But, again, the statistical norm seems to be derived from the fact that dressing as a chicken is first and foremost a violation of a social norm; one *ought* not wear something so absurd to a formal interview.

[30] See for example Hall (2007), Halpern (2016), Halpern and Hitchcock (2015).

cause of *Y* with respect the variables in set ***V**** but not with respect to variables ***V*****. If we are causal realists, we might find this alarming since the selection of the variable sets depend upon how investigators chose to represent the actual state of affairs. The worry is compounded if the modeller choses variables and values on the basis of mind-dependent considerations. Yet this type of relativisation shouldn't worry the realist, says Woodward. Once the combination of variables and values have been fixed, that is, once ***V*** has been specified, the structures of counterfactual dependence therein hold independently of subjective judgements. Whether one variable counterfactually depends upon another is a mind-independent matter of fact; it's in no way determined by how we choose to represent the world or what we have to say about that representation. Thus, the truth or falsity of causal claims produced by a model will similarly be a matter of mind-independent fact. To illustrate consider a concrete example. With respect to a set that includes variables like the assassination of Franz Ferdinand, Austria-Hungary declaring war against Serbia, Russia defending Serbia, and World War I, the assassination of Franz Ferdinand is an actual cause of World War I. If instead one took a more coarse-grained set of variables this may not be true. For example, suppose one focused on the geo-political relations and the expansion of British and French empires, then with respect to variables like these, it would not be true that Franz Ferdinand's assassination caused World War I. Still, this type of relativisation does not entail that the causal facts are subjective, for once the set of variables has been specified, it is a fully objective matter, dependent on the causal structure of the phenomena we are attempting to capture, whether *X* is a cause of *Y* with respect to ***V***. Another way of putting this is to say that whether World War I counterfactually depends on the assassination of Franz Ferdinand or European expansionism is a mind-independent matter. In sum, the idea is that the patterns of counterfactual dependence that make up the foundations of the causal model serve as what Woodward (2003) describes as the "objective core" of interventionism (p. 85). Woodward thus argues that "this sort of relativisation does not introduce an undesirable kind of subjectivism into the characterisation of direct causation in the sense that it makes whether or not *X* is a direct cause of *Y* dependent on the beliefs or psychological state of human investigators" (p. 56).

Several others have taken up a similar line of reasoning. For example, Hitchcock (2007) maintains that even though the models are subjective insofar as they are determined by our selection of variables, the structural equations which represent the causal structure therein are objective. Hitchcock therefore suggests that "we can afford to let judgements of token

causation be infected by pragmatic criteria without giving up on the objectivity of causation generally: objectivity can be retained at the level of token causal structure" (p. 504).

It's a little unclear to me exactly how to categorise such a position. On the one hand, these philosophers accept that the choice of variables is to some extent subjective, and that the causal facts are relative to what model investigators deem as an appropriate representation of the actual situation. On the other hand, they argue that the structural equations or patterns of counterfactual dependence describe objective information about the results of interventions — once a modeler has selected a set of variables to include in the model, the world (not us) determines the truth value of causal claims. Perhaps taken together these claims lend themselves to seeing causation as something partly made-up of subjective and partly objective elements. I'm uncertain to what extent such views would be compatible with causal realism. I suspect the thoroughgoing realist would be hesitant to accept such an account into their camp, given that they see mind-independent features as exclusive determinants of causal facts.

In any case, I don't think we need to settle the question of categorisation just yet. This is because I don't think this strategy retains objectivity to the degree it reports. In particular, I'm sceptical that the structures of counterfactual dependence lying behind the model can secure the kind of objectivity and mind-independence that Woodward and Hitchcock claim that they do. The trouble is that subjective judgements seem to creep in even at the most fundamental level of counterfactual dependence. Counterfactual treatments of causation, one way or another, require us to imagine what would have happened had some event not occurred. To successfully check for counterfactual dependence then, one needs to set a standard for what it means for an event not to occur. My point will be that this standard will be heavily context-sensitive, and plausibly dependent upon mind-dependent notions of normality and abnormality. If this is true, then counterfactual relations will not be able to provide the basis for mind-independency and objectivity after all.

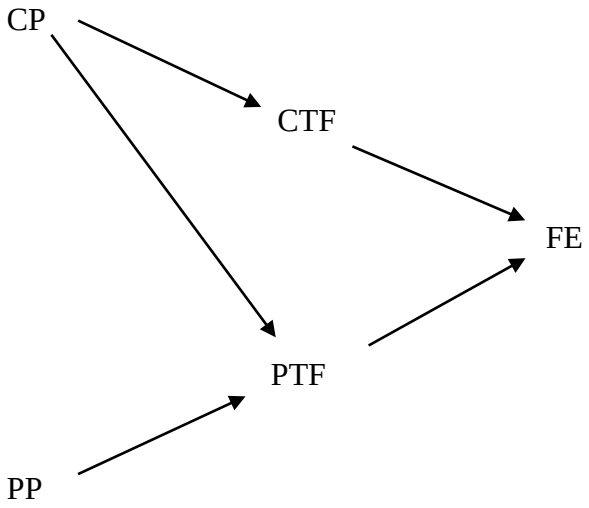
The simplest of examples illustrates the worry. Consider A who puts poison into B's coffee. It is completely natural (and surely correct) to hold that if A had not put the poison into the coffee, B would not have died. But to get here, we must also be supposing that, in the relevant counterfactual scenarios where A does not poison B's coffee, that A (or anyone) is not doing anything else that will also kill B; say, firing a gun at B, forcefully hitting B over the head, hiding B's lifesaving medicine, etc. For if we read the counterfactual this way then it would be false to suppose that if A hadn't poisoned B, then B would not have died. We therefore need a

strategy for reading counterfactuals that permits us to screen-off these erroneous alternatives, and thereby secure the intuitively correct causal verdict.[31]

3.10.1 Late Pre-emption and the Problem of Non-Occurrence

Before considering some such strategies, I first want to demonstrate how this general worry about how to read counterfactuals can produce problems specific to an interventionist framework. To do this, I'll look at the late pre-emption model one final time:

[31] Here I'm raising a worry about how to read counterfactuals that take the form 'If C had not occurred, then E would not have occurred'. These are the kinds of counterfactuals that Lewis (1973) originally employed to determine causal relations. But, of course, interventionism doesn't restrict itself to using only these kinds of counterfactuals. Interventionism's looks for changes in the *values* of variables following combinations of interventions. Sometimes interventions on values involve 'switching off' the variable, in which case causal relations are determined by considering the same kind of antecedent contained in the Lewisian counterfactuals. Other times intervening does not entail 'switching-off' but, say, manipulating the variable's speed. In this case we'd consider a different kind of antecedent; something like 'If X had been moving faster, then…'.
Depending on the specificity of intervention, the worry I'm talking about here can extend to these non-Lewisian antecedents. In the case of manipulating the variable to 'move faster', for instance, it's evident that there's many alternative scenarios we can choose from in which this antecedent is made true. In other cases, however, the worry doesn't hold. For example, manipulating a variable to a different specified colour doesn't allow for much wiggle room in choosing between alternatives – 'If X had been red, then…' elicits a straightforward reading.
To keep things simple, however, I will largely discuss the worries about what it means for an event not to occur, or for a variable to 'switch-off'. Though it's worth bearing in mind that the problem extends to other kind of counterfactuals deployed under an interventionism framework.



Structural Equations:	Actual Values:
CP = 1	CP = 1
PP = 1	PP = 1
CTF = CP	CTF = 1
PTF = PP ∧ ¬ CTF	PTF = 0
FE = CTF ∨ PFT	FE = 1

(Figure 2 in Chapter 3)

In Section 3, I said that Woodward's account is supposed to overcome the problem of late pre-emption because the model is able to differentiate between the cause and the pre-empted back-up. I ended that Section with a note of caution however; I said that though the model seems to get the right causal verdicts, there remain doubts as to whether interventionism really does solve one of the most recalcitrant problems facing counterfactual based theories of causation. One of the reasons why philosophers have doubted the view's ability to solve these cases concerns *the means* by which the model produces its causal claims. The worry is that models like these trade on an ambiguity about what it means for a variable to 'switch-off' or 'not to occur', and it is in virtue of this ambiguity that the model is able to generate intuitively correct claims.

The worry is first articulated by Hall and Paul (2003), then later expanded and developed by Hall (2006). Hall and Paul note that in order to check for actual causation we are to hold the

value of certain variables fixed by keeping them switched off, then we are to calculate the values of the other variables in the model by imagining what would happen to these other values given that we've switched off a variable. But, they argue, sometimes it is unclear exactly what kind of counterfactual scenario we're supposed to envisaging once we've fixed these variables to switch off. In the case that I've been using, the specific problem will arise when we are checking to see if Cat's pouncing is an actual cause of the fly's being eaten. To recap the problem, focus on the directed path {CP-CTF-FE}. When checking for causation along this path, we fix the other direct cause PFT at its redundancy values 1 or 0. Suppose we've set PFT to 0 thereby fixing it by switching it off. We then check for causation by intervening on CP to see if this makes a difference as to whether the fly was eaten, which it does. In order to imagine such a scenario, we're supposing that Cat does not pounce, meaning that Cat does not trap fly, and the fly does not get eaten. The problem is that whilst imaging this scenario we also have to imagine that Possum does pounce, and that Possum does *not* trap the fly. So we have to suppose that Possum does not catch the fly despite the fact that Possum pounces and Cat does not trap the fly. Hall and Paul note that it's completely unclear how such a scenario is supposed to play out. How can it be the case that Possum pounces and Cat doesn't pounce, but nonetheless Possum does not trap the fly? It is only insofar as the model encodes these sorts of ambiguities about how to read the relevant counterfactuals when we switch off a variable that the model is permitted to deliver the correct causal claims.[32]

This argument has sparked quite a thorny debate in causation, and I won't elaborate on it here. The point I wish to make is that Hall and Paul's criticism of interventionism is one manifestation of how problems arise for causal theories when there's a lack of clear standard for what it means for an event not to occur.

3.10.2 Solutions to the Problem of Non-Occurrence

So, what are the relevant kinds of counterfactuals should we be thinking about when considering what it means for an event not to occur? One well-known answer comes from

[32] The problem is compounded once we begin to wonder whether the counterfactual we are being instructed to envisage could even be true according to the model. For if what is to be held fixed are the values of some variables in some apt model, then the details of that model should tell us exactly how to construct the counterfactual scenario. But the structural equations tell us that PTF = 1 whenever Possum pounces and Cat does not pounce. So it's not obvious whether the counterfactual could be true according to the model, and if it can be true there's no structural equation that helps us get to it.

Lewis's (1973) original counterfactual analysis. According to Lewis, the relevant counterfactuals are the ones instantiated in the most similar worlds to the actual world in which the antecedent is satisfied, that is, in the closest worlds where *X* does not occur. So, we ought to hold fixed, as far as possible, the actual facts just up until the point at which *X* occurs and consider the smallest possible alteration to the actual facts that would entail the non-occurrence of *X*. When we imagine the non-occurrence of A's poisoning, we should not look to worlds where the poisoning is replaced by A hiding B's lifesaving medicine or A hitting B over the head, for such alterations do not constitute the closest worlds in which the antecedent is made true. However, the smallest alteration strategy doesn't guarantee that we'll screen-off the wrong sorts of counterfactuals. Sometimes the non-occurrence of *X* can be brought about by a very small alteration, and in such cases, this can lead to very similar outcome, thus turning what we intuitively think of as a cause into a non-cause. If A had put a little less or a little more poison into B's coffee, the actual event of A's poisoning would not have occurred, and yet, B would have died all the same. Consequently, A's poisoning would not be a cause of B's death. This is obviously wrong. The problem has been noticed before by, for example, Lewis himself. Here's his gloss of the problem and a preliminary solution:

> What is the closest way to actuality for *C* not to occur? It is for *C* to be replaced by a very similar event, one that is almost but not quite *C*, one that is just barely over the border between versions of *C* itself and its nearest alternatives. But if *C* is taken to be fairly fragile, then, if almost-*C* occurred instead of *C*, very likely the effects of almost-*C* would-be almost the same as the effects of *C*. So our causal counterfactual will not mean what we thought it meant, and it may well not have the truth value we thought it had. When asked to suppose counterfactually that *C* does not occur, we do not really look for the very closest possible world where *C*s conditions of occurrence are not quite satisfied. Rather, we imagine that *C* is completely and cleanly excised from history, leaving behind no fragment or approximation of itself. (Lewis 2004, p. 90)

Replacing the small alteration strategy with the excision strategy now gets us the right answer with regards to A and B: if we consider the closest world where A's poisoning is completely and cleanly excised from history leaving no fragment or approximation of itself, then the counterfactual comes out as true — no poisoning, no death. However, as some philosophers have pointed out, the excision strategy generates more problems than it solves. The complete

excision of an event will presumably consist in plucking the event out of history which will leave behind a mysterious void where the actual event was located. And whilst the creation of a void is odd in itself, the real trouble consists in the causal consequences of creating such a void. Neil McDonnell (2019), for instance, argues that "the sudden appearance of a void will have an enormous impact on every surrounding region, and thus the complete excision of any (ordinary) event, will perturb all of its surroundings. Thus, by the lights of a counterfactual analysis, every event would be a cause of this surrounding (p. 275)." Replacing the actual event with a void would produce many more diverse effects than in the actual world, far more effects than we would like to endorse. As a result, the excision strategy is "simply too blunt a test for causation by the lights of our common sense judgements" (2019, p. 275).

If the excision strategy and the small alteration strategy are unsuccessful, is there another method one can employ for setting the standard for non-occurrence? Yes, but before offering up my suggestion, let me diagnose exactly what goes wrong when erroneous counterfactuals are envisaged to determine causal relations. The trouble arises when we imagine the non-occurrence of the event by replacing the occurrence of that event with a similar iteration of it, or by replacing it with some distinct event that nonetheless produces the effect. A small alteration in the event which merely manifest an iteration of the event will sometimes produce the effect i.e., putting a little less poison in the coffee will still cause death. The same is true when we replace the event with a distinct event that can nonetheless produce the effect i.e., shooting rather than poising will still cause death. Hence, the counterfactuals we want to screen-off are those where we replace the actual event with something *else* that would produce the effect. I suggest that applying a distinction between normal and abnormal states can assist us with this. Conceive of the characteristic way things would usually develop as normal states and conceive of abnormal states as those things which disrupt the normal states. The standard for non-occurrence is one that returns the abnormal to the normal. So, imagining what would have happened had *X* not occurred entails replacing *X* — an abnormality — with something normal. This idea builds upon Hart and Honoré's (1959) claim discussed in Chapter 2 that causation is an intervention that makes a difference to the way things would normally develop. To illustrate, the right type of counterfactual to invoke when checking for A's causal involvement in B's death, is one where A returns to a normal state of, say, watching TV or standing idly whilst B makes his coffee. Envisaging a scenario where things develop as we would normally expect them to allows us to screen-off those erroneous alternatives whereby something else happens to make the effect occur. This allows us to get to the claim that A's

poisoning caused B's death. As a further illustration consider the counterfactual 'If Suzy had not thrown the rock, the window would not have shattered'. To get the verdict that Suzy's rock throwing caused the window to shatter, we have to imagine a scenario in which Suzy is not throwing her rock and that she nor anybody else is doing something that would shatter the window. This can be achieved by replacing Suzy's throw with something that represents a normal state, say, Suzy standing by or walking past the window. Hall and Paul (2013) briefly gesture towards a similar strategy for supplying a non-occurrence standard. Indeed, they note that in regards to the problem of non-occurrence "we strongly suspect that any ontological reduction of causation that makes use of counterfactuals will need to deploy some distinction between default states and deviations thereof" (p. 36). I anticipate that as with the other constraints and criteria I've considered in this Chapter, that the relevant kind of normality being appealed to when returning *X* to its normal state will be one that sometimes appeals to norms that are determined by our minds and perspectives. I won't thoroughly argue for this strategy here. But let me say that understanding what it is for an event not to occur in a way that generates intuitively correct causal claims requires that we screen-off counterfactuals in which we imagine something else magically occurring at the relevant time which would produce the effect. This looks to me to be the same as saying that the relevant counterfactuals are the ones where things unfold as they would normally do.

Bringing this discussion back to interventionism, if features such as normality and abnormality (in particular a mind-dependent rendering of normality and abnormality) are decisive in determining the truth value of counterfactual conditionals, then the patterns of counterfactual dependence which provide the scaffolding for causal models are not necessarily mind-independent in way Woodward supposes. In light of this, I'm sceptical that one can locate complete mind-independency even at the most fundamental level of counterfactual dependence.

3.11 Conclusion

In this Chapter my chief aim has been to analyse whether interventionism is compatible with a realist conception of causation. I began the analysis by arguing that the apparatus through which interventionism determines causal relations — causal modelling — is successful to the extent that the models appropriately represent the causal story. I then outlined two criteria that are typically invoked to determine the aptness of a model — the stability criterion and the serious possibility criterion. These criteria are presumed to be realist-friendly in the sense that

they do not obviously appeal to normative considerations. However, I've tried to show the criteria are essentially bound up in or need to be supplemented with an appeal to what's normal and abnormal. In the case of stability, I suggested that in order to measure the stability of causal relationships within a model one must first differentiate between causes and background conditions, and second make certain relevant manipulations on these background conditions. Both of these steps, I suggested, require an appeal to what's normal and abnormal. In the case of the serious possibility criterion, I argued that the criterion is helpful in excluding certain variables and values from being represented in the model but it's not especially helpful in guiding us as to what to include in the model. I ended by arguing that considering what it is for something to be normal or abnormal in the context can fill this lacuna. Crucially, throughout presenting my argument I've demonstrated that the relevant concepts of normality and abnormality we're dealing with when it comes to applying the stability and serious possibility criteria appeal to norms that are mind-dependent.

As noted, the idea that interventionism needs to rely on considerations about what's normal has been argued for before by, for example, Halpern (2016), Halpern and Hitchcock (2015) and Hall (2007). These philosophers do so by arguing that there ought to be additional constraints on what makes a model apt, and that these constraints ought to explicitly appeal to normative notions. I take my argument in this Chapter to therefore be contributing to the general argument that interventionism must appeal to norms for its success, but for different reasons than have already been given. Instead of arguing that one needs additional normative constraints on model construction, I've argued, at least in the case of stability, that the existing supposedly realist constraints bottom-out into normative constraints.

So where does this leave interventionism in terms of the taxonomy I created in the previous Chapter? To recap, I presented three meta-causal approaches to causation: pure realism, restricted normativism and unrestricted normativism. The first two approaches are compatible with a realist conception of causation since they determine causal relations by calling exclusively upon mind-independent features of the world. By contrast, unrestricted normativism is not compatible with a realist conception of causation because it determines causal relations by invoking both mind-independent and mind-dependent features of the world. I've argued that both mind-independent and mind-dependent norms are integral for building the correct models, and since the models determine what causal relations there are, a successful rendering of interventionism is one that sits in the unrestricted normativism camp, and therefore is incompatible with causal realism.

CHAPTER 4

Mapping the Way

In the previous two chapters I explored the ways in which facts about morality, and normative considerations more widely, determine causal facts of the kind '*c* caused *e'*. In Chapter 2 I began by looking at two 'meta-causal' views: normativism and realism. I described meta-causal views as those which do not aim to specify the necessary and sufficient conditions that make claims like '*c* caused *e*' true, but rather aim to delineate the wider systemic features of causation, the details of which bottom-out in various and diverse ways. Having outlined what normativism and realism take the systemic features of causation to be, I then argued, contra to the prominent philosophical discourse, that these two views were to some extent compatible. I showed that causal realism can appeal to a restricted set of normative considerations to determine causal facts. This discussion left three kinds of meta-causal views on the table: "pure realism", "restricted normativism", and "unrestricted normativism". Pure realism and restricted normativism are compatible with a realist conception of causation, whilst unrestricted normativism is not due to the fact it draws upon mind-dependent norms to determine causal relations.

In Chapter 3, I went onto to analyse whether one particular account of causation is compatible with causal realism. The account I focused on was James Woodward's (2003) interventionism. My aim in this Chapter was to provide reasons in favour of thinking that a successful interventionist theory is incompatible with causal realism. My argument turned on the fact that the apparatus through which interventionists determine the causal facts — causal models — depend for their success on mind-dependent notions of normality and abnormality. As a result, this put interventionism in the unrestricted normativist camp.

In the next two Chapters, I move to explore the other direction of the dependency relationship between causation and morality. Namely, I examine the ways in which causal facts determine moral facts. In Chapter 5, I argue that:

> Whether an agent is morally responsible for an outcome partly depends upon whether she was a cause of that outcome.

And in Chapter 6, I argue that:

> Whether a right action is morally praiseworthy partly depends upon the extent to which the agent's motive would cause her to act rightly in alternative scenarios.

In both Chapters 5 and 6 then, I argue that the kind of causal facts that determine certain claims about morality include facts about when one thing causes another i.e., facts about '*c* caused *e*'. Now given my focus on interventionism in Chapter 3, the reader might expect me to adopt an interventionist framework to establish such facts. However, I chose a different strategy; instead, I adopt the simple counterfactual analysis. This might be surprising given that in the Introduction and Chapter 3, I framed the simple counterfactual analysis as a theory which has now been supplanted by interventionism. The aim of this short Chapter is to explain my motivations behind this choice.

There are two chief reasons why I've chosen to focus on the simple counterfactual analysis over interventionism. The first reason concerns simplicity. Although both accounts take their departure from the same basic thought that *'c* caused *e*' is true when *c* makes a difference to *e*, the two views formalise this idea through very different frameworks. The simple counterfactual analysis establishes whether *c* caused *e* by merely checking for counterfactual dependence between the effect event and the cause event. Interventionism, however captures the difference making relation through a much more complex conceptual framework. Firstly, the truth value of '*c* caused *e*' statements are established by reference to several distinct causal concepts — including direct cause, directed causal paths and redundancy range — each of these concepts are given their own precise definition. Secondly, the account determines whether actual causal relations exist in the world by employing a formal apparatus of directed graphs and structural equations that together create causal models. Moreover, such models are subject to independent criteria that aim to adjudicate for their aptness in representing the actual causal structure. And once these models are created, causation is tested for by imagining a possible combination of interventions on the values of the variables within the model. Evidently, interventionism draws on many more conceptual resources, and is therefore a more complex theory of causation than the simple counterfactual analysis.[33]

[33] I appreciate that the simple counterfactual analysis comes with conceptual baggage too. This is especially true if one is adopting the Lewisian/Stalnaker possible world semantic framework to determine the truth value of counterfactual conditionals. Still, once these metaphysical frameworks have been settled establishing the truth value of causal claims is far more taxing under the interventionist approach than the simple counterfactual approach.

However, presumably simplicity *alone* is not sufficient to motivate the adoption of the simple counterfactual analysis over interventionism. After all, I argued that interventionism is regarded as more successful than the simple counterfactual test, and one might think that we should adopt the more successful approach even if this results in a more complicated theory. To this let me say that in terms of what I need a theory of causation to do in the upcoming Chapters, interventionism and the counterfactual analysis are on par. By this I mean that I can achieve what I set out to in the following Chapters either by using the interventionist framework or by using the simple counterfactual test. The reasons these two approaches are equally successful in achieving my aims is twofold.

Firstly, in Chapter 6 I provide a positive proposal for determining the truth value of claims about moral praiseworthiness which appeals to facts about when *c* caused *e*. To motivate my proposal, I consider cases which have fairly straightforward causal structures. This was intentional so as to not pull focus from the target phenomena — moral praise. In light of the fact that the considered cases are relatively straightforward causally speaking, the simple counterfactual analysis gets the intuitively correct verdict when it comes to identifying when one thing causes another. It therefore achieves the objectives I need it to when it comes to delivering my positive proposal. I anticipate that interventionism will be equally successful in underpinning my proposal by way of delivering the intuitively correct verdicts in such cases, given that interventionism is an extension of the simple counterfactual analysis. Indeed, as we saw in the previous Chapter, interventionism can revert to checking for causation using the simple counterfactual test simply by 'switching off' a variable's value. So, in terms of what I need an account of causation to do in Chapter 6 both views will do an equally good job.[34] Given that we achieve the same results using either the simple counterfactual analysis or interventionism, I have opted for the simple counterfactual test so as to not unnecessarily overcomplicate matters.

Secondly, in Chapter 5, I provide a positive proposal for determining the truth value for claims about moral responsibility which also appeals to facts about when *c* caused *e*. Unlike Chapter 6, the cases I consider in this Chapter have complex or unconventional causal structures,

[34] It's perhaps worth pointing out that if one did want to consider more complex causal structures in cases of moral praise, in particular structures regarding redundant causation, then I suspect that interventionism will be more successful than the simple counterfactual test at getting to the correct causal facts. But I imagine that such unconventional causal structures are few and far between when it comes to agents acting in a praiseworthy manner, and to consider such cases would probably needlessly complicate things.

specifically, they involve cases with omissions. Given the complexity of these cases, I argue that the simple counterfactual analysis does have the resources to deliver the intuitively correct causal verdicts, I therefore go on to supplement the counterfactual analysis with a condition that appeals to different types of causal facts; namely facts about causal stability. At this juncture, one might wonder why I opted for supplementing the view with an appeal to causal stability, rather than looking to see if the more conceptually rich interventionism would be sufficient on its own to deliver the intuitively right causal verdicts for cases involving omissions. The reasons I don't go down this route is because interventionism, despite having more resources to hand, fails to deliver the right causal claims in the same way the simple counterfactual analysis does. (I outline the precise reasons why interventionism fails with these sorts of cases in footnote no. 39). So, whilst the simple counterfactual analysis is just as successful as interventionism at underpinning my positive proposal in Chapter 6, it is equally unsuccessful when it comes to underpinning my positive proposal Chapter 5. Given that they're equivalent in this regard, I focus on the counterfactual analysis because it has the edge in terms of simplicity

The second reason I dispense with interventionism going forward concerns the wider context of the philosophical debate. As mentioned, the central focus in the next two Chapters is how we determine the truth value of certain *moral* claims. Given that the aim is to analyse the determiners of moral facts, I approach the analysis by primarily engaging with the philosophical literature in ethics. Most of the authors in this literature who have written about causation's effect on moral assessments have done so with reference to the simple counterfactual analysis. That is, they have regarded causal claims of the kind '*c* caused *e*' to be determined by the simple counterfactual test (I cite some evidence for this in the next chapter). In this way, talk of interventionism would fail to engage with the discussion in a manner that is currently understood, and would presumably therefore make a smaller or less significant contribution to the philosophical literature. So, I leave interventionism here. Chiefly for reasons having to do with simplicity and for reasons having to do with engaging in the contemporary philosophical debate.

CHAPTER 5

Moral and Causal Responsibility for Omissions

5.1 Introduction

According to a widely accepted claim in ethics, to be morally responsible for an outcome requires that one is also causally responsible for that outcome. Let us call this condition on moral responsibility the Causal Requirement:

> **Causal Requirement (CR):** An agent is morally responsible for a particular outcome only when she is causally responsible for that outcome.

According to another widely held claim, agents can be morally responsible for outcomes that arise from omissions. Let us call this idea Moral Responsibility for Omissions:

> **Moral Responsibility for Omissions (MRO):** Agents can be morally responsible for outcomes that result from failures to act.

CR and MRO are immediately intuitive. But notice that if one were to endorse *both* of these claims then one faces a pivotal challenge. If one accepts that an agent is morally responsible for an outcome only when they are causally responsible for that outcome *and* that agents can be morally responsible for outcomes that result from omissions, then one must accommodate for the idea that agents can be causally responsible for outcomes through their omissions. So, the challenge is to provide an account of *omissive causal responsibility*.

In this Chapter, I examine various ways one might meet this challenge. The first strategy I look at assigns omissive causal responsibility on the basis of the simple counterfactual test for causation. Though there has been very little said about omissive causal responsibility (and causal responsibility more widely) in the philosophical literature, those who do make reference to it generally defer to what the dominant theory of causation — the counterfactual analysis — has to say about it. According to this approach, an agent is causally responsible for some outcome just when the counterfactual analysis says that the agent's omission is a cause of the outcome. In other words, the counterfactual test for causation suffices as a test for omissive causal responsibility.

I argue that the counterfactual test is a poor candidate for a theory of omissive causal responsibility. The crux of the problem lies in the fact that the counterfactual test renders nearly every omission a cause of some outcome. In terms of omissive causal responsibility, this means that every one of our literally countless nondoings makes us causally responsible for some outcome. This is problematic, I argue, because not every one of our countless nondoings makes us causally responsible, only those nondoings which genuinely reflective of a person's agency can make one causally responsible.

I next consider one way to amend the counterfactual test with a view to limiting its attribution of omissive causal responsibility to those agency-indicating omissions. The amendment I have in mind supplements the test with an appeal to normative considerations. According to this approach, from the countless number of nondoings happening at every moment, only those nondoings that violate a norm can make one causally responsibility. Absent some such norm violation, the omission is said to cause nothing at all.

Whilst I think this approach gets closer to supplying an account omissive causal responsibility, I will argue that appealing to normative considerations generates unique problems. Firstly, it leaves some relevant omissions out of the picture. For there are occasions when we intuitively take someone to be a cause of some outcome, yet their omission does not violate a norm. And secondly, there seems to be something objectionable about basing attributions of omissive causal responsibility on norms when the norms themselves are pernicious.

After raising doubts over these two approaches, I will set about undertaking the primary ambition of this Chapter, which is to provide my own account of omissive causal responsibility. Like the normative approach, my view supplements the counterfactual analysis with a view to restricting attributions of omissive causal responsibility to the relevant agency-indicating omissions. But unlike the normative approach, I do not appeal to normative considerations to do this. Rather, I appeal to the notion of causal stability. I'll suggest that an agent is omissive causally responsible for an outcome when the causal connection between her omission and the outcome is a relatively stable one. Whereas an agent is not omissive causally responsible for an outcome when the causal connection between her omission and the outcome is a relatively unstable one. Thus, I'll be arguing that being causally responsible for an outcome as a result of an omission requires causal stability. The idea being that those omissions which are causally stable are also the ones which are reflective of a person's agency and can therefore form the basis for attributions of omissive causal responsibility. I previously introduced the notion of

causal stability in Chapter 3, however the notion of stability I outlined then is slightly different from the one I employ in this Chapter. Most notably, the notion invoked in this Chapter does not appeal to normative considerations. I outline my delineation of causal stability and explain why it different from Chapter 3's delineation in Section 10.1.

To clarify, my aim in this Chapter is not to defend CR and MRO per se. Rather, my aim is to offer one plausible method for advocates of CR and MRO to establish omissive causal responsibility in a way that allows them to ground judgements of moral responsibility.

Roadmap: I begin by offering some clarifications about what sort of project is being undertaken here. In Sections 3 and 4 I outline CR and MRO in more detail. In Section 5, I set forth the challenge to establish a theory of omissive causal responsibility for philosophers who endorse both CR and MRO. In Section 6, I evaluate how one might establish omissive causal responsibility by using the counterfactual test for causation. In Sections 7 and 8, I fix the target and set the success conditions for a theory of omissive causal responsibility. In Section 9, I investigate and reject supplementing the counterfactual test with a normative framework to test for omissive causal responsibility. In Section 10, I explicate and defend my own view of omissive causal responsibility.

5.2 Clarifications

First, let me clarify the scope of the project. It's natural to draw a distinction between omissions themselves and how things turn out in the world as a result of those omissions. Let's call this last category of things outcomes. The concept of moral responsibility applies quite broadly; in particular, we can hold agents morally responsible both for their omissions and for the outcomes of those omissions. Causal responsibility, on the other hand, applies most principally to outcomes. As a result, I will be focusing on how we assign moral and causal responsibility for outcomes.

Secondly, causal responsibility is typically taken to be a necessary condition on moral responsibility, but it certainly isn't sufficient — an agent can be causally responsible for some outcome without being morally responsible for that outcome. Imagine that at a party I am pushed by one of the other guests, I lose my balance, and I spill red wine on the carpet. Even if I'm causally responsible for the wine stain I am clearly not to be blamed for it. To be blameworthy (or praiseworthy) one's act must be a voluntary one. In addition to voluntariness and causality, many argue that agents must also be able to reasonably foresee the consequences of their action or inaction in order to be held morally responsible. Imagine that I intentionally

spilled my wine. To be blameworthy it must be the case that I could reasonably foresee that the spill would cause damage to the carpet. There may be further conditions on moral responsibility, but what's important to bear in mind for the purposes of this Chapter, is that being causally responsible does not yet make one morally responsible, and therefore, establishing causal responsibility should not be equated with establishing moral responsibility.

Finally, a brief note about terminology. There are various ways we can describe an omission; we might say that when someone omits to act they "refrain", "fail", "forget", "withhold" or simply "do nothing". Some authors ascribe specific meaning to these different terms. For instance, Joel Feinberg (1987) sees omissions and refraining as implying two separate though related phenomena. He says of someone who doesn't water the plant that "though perhaps he did *refrain* from doing so, having briefly considered the matter and then decided against it", the agent hasn't omitted to water them unless there was a special duty or requirement that he do it (p. 161). For my purposes, however, there's no need to demarcate different meanings for these various terms, so I will refer to them all interchangeably.

5.3 The Causal Requirement

As mentioned in the introduction, I don't intend to defend either CR or MRO in this Chapter. Nevertheless, I will begin by briefly motivating each of these claims, in the hope that doing so will, at the very least, provide prima facie reasons to suggest that developing an account of omissive causal responsibility is a challenge worth pursuing. I'll start with the idea that moral responsibility requires causal responsibility.

> **Causal Requirement (CR):** An agent is morally responsible for a particular outcome only when she is causally responsible for that outcome.

CR has intuitive appeal. Imagine that a robbery takes place in a distant part of the world you've never visited or even heard of before. Should we hold you morally responsible for the robbery? Of course not. You cannot be held morally responsible for something if your actions or inactions bear absolutely no relation to it. To be morally responsible for something we need to be hooked-up to it in some way, which is to say that our actions or omissions must be connected to the event for which we want to attribute moral responsibility. One plausible way to ground this connection is by way of causation; our actions or inactions connects us with outcomes by means of what they cause. The natural thought then is that to be morally responsible for an outcome our actions must have caused that outcome to occur.

CR enjoys widespread support. Many philosophers either explicitly or implicitly endorse the claim. For instance, Gideon Rosen (2004) argues that to be culpable for a morally bad action the action must have derived from a distinctive sort of "inculpating causal history" on behalf of the agent (p. 309). Susan Wolf (2015) seems to advocate for a causal requirement on responsibility insofar as a causal connection is required for us to make judgements about when one is "deeply responsible" for some outcome (p. 130). Furthermore, writers like Scanlon (2008) appear to take causation as a requirement for blame insofar as blame constitutes one judging that another has caused an impairment in an interpersonal relationship.

Others are more explicit in their endorsement of CR. For example, Joel Feinberg (1970) proposes a theory of moral responsibility that contains a causal condition component, and Julia Driver (2008) argues that '[i]f an agent A is morally responsible for an event e, then A performed an action or an omission that caused e' (p. 423). Furthermore, Marina Oshana (1997) contends that "when we say a person is morally responsible for something, we are essentially saying that the person did or caused some act (or exhibited some character trait) for which it is fitting that she give an account" (p. 77). Similarly, Matthew Braham and Martin van Hees (2012) argue for a "causal relevancy condition" on moral responsibility which entails that "there should be a causal relation between the action of the agent and the resultant state of affairs" (p. 605).

Elsewhere in the philosophy of law, Michael Moore forcefully shows in *Causation and Responsibility* (2009) that causation is an explicit and essential element in most doctrines of legal liability. In fact, Moore suggests that "it is plausible to think that all liability doctrines in criminal and tort laws require that a defendant's act cause something" (p. 19, original emphasis). Although this Chapter is not about legal liability, it is significant that causation is taken to be a requirement in criminal and tort laws since judgements in these domains are grounded upon, amongst other things, facts about moral responsibility.

As intuitive as CR may be, some have argued that one may be responsible for an outcome without being causally responsible for it. For instance, Sartorio (2004) designs a pair of elegant counterexamples to CR. In the first case A_1 and A_2 independently and simultaneously fail to depress two buttons which, had they been pressed, would have prevented an explosion. In the second case, A_1 fails to press the button as before, but the second button is not depressed because of a stuck safety mechanism. Intuitively, A_1 is morally responsible for the explosion in the first case. Further, Sartorio states that intuitively A_1 is not causally responsible for the

explosion in the second case. Since omissions are metaphysically equivalent, it follows that A_1's failure to depress the button in the first case doesn't make her causally responsible for the explosion. Consequently, A_1 is morally responsible for the explosion in the first case without having caused it (p. 318).

Although an interesting pair of cases, I have a couple of concerns with Sartorio's argument. Most relevantly, I don't share the intuition that A_1 is not causally responsible for the explosion in the second case. Granted, A_1's omission is not a 'singular' cause thanks to the faulty safety mechanism, still it seems to me that A_1's failure makes her a causal factor, perhaps a 'joint' causal factor. Indeed, when talking about causation in the context of moral responsibility attributions we don't seem to have in mind a narrow concept of causation that equates to only singular causes, but a more liberal notion that includes the variety of ways our actions or omissions can involve themselves in the causal structure leading to an outcome.

5.4 Moral Responsibility for Omissions

Next consider the idea that we can be morally responsible for something when we do not act.

> **Moral Responsibility for Omissions (MRO):** Agents can be morally responsible for outcomes that result from failures to act.

Like CR, MRO has immediate intuitive appeal. We regularly praise people for the good that arises from their not doing things. For instance, we might think Cody is praiseworthy for deescalating a conflict in virtue of the fact that he does not retaliate when he is provoked. We also blame people for not doing things. For instance, we might hold Kate blameworthy for her grandmother's hurt feelings on account of the fact that Kate forgot to call her grandmother to wish her a happy birthday. In praising or blaming these people we necessarily take them to be morally responsible for the outcomes of their omissions.

MRO is widely accepted by those working on ethics and agency. For instance, Derk Pereboom (2015) argues that "just as we can be morally responsible for decisions, it seems that we can be morally responsible for failures to decide. And as we can be responsible for the outcomes of our decision to act, we can also be responsible for the outcomes of decisions not to act, and the outcomes of failures to decide" (p. 191). Similarly Dana Kay Nelkin and Samuel Rickless (2017) state that "common-sense morality holds many unwitting omitters are morally responsible for their omissions (*and for the consequences thereof*)" (p. 106, my emphasis).

Moreover, in his book *Who Knew,* George Sher (2009) considers and proposes to justify our practices of holding people responsible for outcomes of omissions in cases where the epistemic conditions on responsibility appears not to be satisfied. Randolph Clarke (2014) proposes a similar framework to Sher for holding people responsible for outcomes in such cases (pp. 171-172). Elsewhere, Joseph Metz (2021) takes it for granted that agents can be morally responsible for the outcomes through their omissions, and consequently goes on to provide an account of collective responsibility for outcomes that occur through omissions.

There endorsements of MRO are well motivated. In some instances the denial of MRO would be extremely implausible. To illustrate, consider an example given by Douglas Husak:

> Susan is driving reasonably along a wide lane in a suburban neighbourhood. Since she is moving a bit downhill, her foot is off the accelerator. Her hands are on the steering wheel when she first observes a pedestrian (Megan) crossing the street outside of a crosswalk approximately 100 feet in front of her. She knows she could easily move the wheel to avoid hitting Megan without veering outside of her lane. Instead, she deliberately remains motionless and allows her car to stay on its present course. Her car hits and kills Megan a second or two after she first noticed her. (2017, p. 166)

Husak argues that Susan is liable for Megan's death.[35] Specifically, he argues that Susan's failure to move her car steering wheel makes her liable for Megan's death, and that to argue that someone or something else is liable would be highly counterintuitive. If we want to make sense of Husak's judgements (and I think we ought to), then we need to accept MRO, for if we don't, we would be unable to claim that Susan is morally responsible for Megan's death *because* Susan failed to move the car's steering wheel.

Although I take MRO to be a fairly uncontroversial claim, omissions as entities in themselves are controversial. When Cody decides not to retaliate or when Kate forgets to call her grandmother, what is it that they're doing? The philosophical literature across agency, metaphysics and ethics offers a number of views on what omitting to act amounts to. Some authors, like Hart and Honoré (1985) and Jonathan Schaffer (2004), have taken omissions to reduce to quite ordinary positive actions. These philosophers believe that omissions are merely redescriptions of bone fide positive events, so that expressions like "her failure to sit" refer to

[35] Husak uses this case in a discussion of legal liability as opposed to moral liability.

identical positive actions like "her continuing to stand". Some agree that omissions are actions, but not redescriptions of positive acts, rather they're taken to be distinctively negative actions. For instance, Douglas Walton (1980) maintains that "omissions, refrainings, [and] forebearances" are kinds of "negative actions" (p. 319). Still, others argue that omission do not reduce to actions in any sense but are instead exactly the absence of actions; when an agent omits to act, there's nothing at all that is the agent's omission. Clarke (2014) for example holds than in some cases when there is a failure to act, there is nothing that is the omission.[36] A fourth response, defended by Sara Bernstein (2014), characterises omissions as a tripartite metaphysical entity comprised of an event at a possible world, an event in the actual world, and a counterpart relation between the two. On this view, when one fails to keep a promise that omission comprises of an actual event (whatever one was doing in the actual world instead of keeping the promise) and one's actually keeping the promise at a possible world.

I do not take a stand with regards to the ontology of omissions here. But let me say that omissions are sometimes distinguished into roughly two categories — intentional omissions and unintentional omissions.[37] When Kate forgets to call her grandmother she omits unintentionally, but when Cody consciously refrains from retaliating he intentionally omits to act. What I have to say about omissive causal responsibility applies to both intentional and unintentional omissions.

5.5 The Challenge: Omissive Causal Responsibility

Now, if one accepts that an agent is morally responsible for an outcome only when they are causally responsible for that outcome (CR) *and* that agents can be morally responsible for outcomes that result from omissions (MRO), then one must accommodate for the idea that agents can be causally responsible for outcomes through their omissions. So, the challenge is to provide an account of omissive causal responsibility.

Let me now recap what's to come. In the rest of this Chapter, I explore various ways one might meet this challenge. And, as a reminder, all of these approaches will be consistent with any of

[36] Notably, however, Clarke believes that not all omissions are nothings; he sees some omissions as identical to actions. For example, there are actions of refraining, such as placing one hand over another to refrain from tapping one's hand, that count as an omission according to Clarke.

[37] Distinguishing between intentional and unintentional omissions has become convention in the literature since Patricia Smith's paper 'Contemplating Failure: The Importance of Unconscious Omission' (1990). In the paper Smith urges future work on omissions to note an important distinction between simply not acting as a result of an unconscious failing from consciously refraining from acting.

the views about the ontology of omissions mentioned above. The first strategy I look at assigns omissive causal responsibility on the basis of the counterfactual test for causation. I'll argue that we should not adopt this strategy because the test does not pick out the kind of omissions that have the power to make one causally responsible i.e., agency-indicating omissions. I then go onto to consider one way we might amend the counterfactual with a view to appropriately limiting its attribution of omissive causal responsibility to those relevant omissions. The amendment I have in mind is defended by Sarah McGrath (2005), who supplements the counterfactual test with an appeal to normative considerations. Whilst I think the normative approach gets closer to an account of omissive causal responsibility, I ultimately argue that this view also fails to pick out the relevant agency-indicating omissions.

Later in the Chapter, I will turn to my main aim: to provide my own account of omissive causal responsibility. Like McGrath's account, my account supplements the counterfactual analysis with a view to restricting attributions of omissive causal responsibility to the relevant type of omissions. But rather than appeal to norms, my account appeals to the notion of causal stability. In short, I argue that an agent is causally responsible for an outcome through her failure to act when the causal connection between the outcome and her omission is a relatively stable one. It will become clear what a fully-fledged account will look like later, for now let us start to look at how one might establish omissive causal responsibility via the counterfactual analysis of causation.

5.6 The Counterfactual Analysis

It's fair to say that the counterfactual analysis of causation championed by David Lewis (1973) has been the most influential approach to causation in metaphysics. The simplest treatment takes as it point of departure from the idea that causes make a difference to their effects, and that this difference-making relation should be understood in terms of counterfactual dependence between the effect, *e*, and its cause, *c*. I outlined the bones of the analysis in previous chapters, now I'll fill out the view in a little more detail since it occupies a more central role in the discussion. Following Lewis, the occurrence of *e* is counterfactually dependent on the occurrence of *c* if and only if the following two counterfactuals are true:

$O(c) \mathrel{\square\!\!\rightarrow} O(e)$

$\neg O(c) \mathrel{\square\!\!\rightarrow} \neg O(e)$

Here, *O(c)* represents the proposition that *c* occurs, the □→ symbol represents the counterfactual conditional, and the ¬ symbol represents negation. So, in English, *O(c)* □→ *O(e)* corresponds to: if *c* were to occur then *e* would occur. And ¬*O(c)* □→ ¬*O(e)* corresponds to: if *c* had not occurred then *e* would not have occurred. The first of these counterfactuals will be true whenever *c* and *e* occur (at least according to Lewis's semantics for counterfactuals). When this is the case, the truth conditions for causation can be simplified somewhat. For when *c* and *e* occur, *e* depends causally on *c* iff, if *c* had not been then *e* would not have been. This is just the second counterfactual. For this reason, most of the attention in the causation literature has focused on Lewis's second counterfactual. I mention the first counterfactual here because it will do important work in capturing aspects of causation which I take to be relevant to omissive causal responsibility, namely causal stability. I discuss exactly what work it does in this regard in Section 10.

Focusing on the second of these counterfactuals for now then, the simplest version of the view states that some event *c*, is a cause of another event, *e*, when there is counterfactual dependence between *e* and *c*, such that if *c* had not occurred, then *e* would not have occurred. To illustrate, suppose that Suzy throws a rock at a window and the window shatters. Suzy is a cause of the shattered window, according to the counterfactual test, because it is true that had Suzy not thrown the rock, the window would not have shattered; there is counterfactual dependence between the shattered window and Suzy's throwing of the rock.

Typically, ethicists who advocate for CR don't go as far as to remark on what they take causal responsibility for actions to entail. Still, of those handful of philosophers who do try to formalise causal responsibility they do so largely by employing the simple counterfactual test.[38] According to this strategy, if an agent's action meets the conditions for causation under the counterfactual analysis this is sufficient to grant causal responsibility; causation equates to causal responsibility. Applied to Suzy, the strategy allows us to hold Suzy causally responsible for the shattered window since her action is a cause of the shattered window. This type of approach to determining causal responsibility enjoys considerable dominance within criminal and tort law. As Michael Moore (2009), notes "whether the law calls it the 'but for' test, the *'sine qua non*' test, the 'necessary condition' test, or something else, it is plain as daylight that what is meant is the identification of the natural relation of causation with counterfactual

[38] Driver (2008), Petersson (2013).

dependence" (p. 371). When it comes to law, Moore (2009) notes that statements of the form '*C* caused *E*' are taken to mean '*E* counterfactually depends upon *C*' (p. 371).

To see how the approach works when it comes to assigning omissive causal responsibility, let's consider Husak's story of Susan and Megan again. Here is a truncated version I'll call **Driving**:

> Susan is driving in a suburban neighbourhood. Her hands are on the steering wheel when she observes a pedestrian (Megan) crossing the street approximately 100 feet in front of her. She knows she could easily move the wheel to avoid hitting Megan without veering outside of her lane. Instead, she deliberately remains motionless. Her car hits and kills Megan.

Let's suppose that Susan is morally responsible for Megan's death, and that her failure to move the steering wheel is the thing that makes her morally responsible for the death. Is the counterfactual analysis able to ground this verdict by attributing omissive causal responsibility to Susan? On the face of it, yes. To see this, first we need to check for counterfactual dependence between Susan's failure to turn the wheel and Megan's death. Testing for counterfactual dependence between *events* involves that we consider whether *e* would have occurred, had *c* not occurred. When *c* is an omission, we should consider what would have happened to the effect had the non-occurrence of the omission occurred. In other words, we ask whether *e* would have occurred, had *not c* not occurred. An equivalent and simpler way of asking whether *e* would have occurred had *not c* not occurred, is to ask whether *e* would have occurred, had *c* actually *occurred.* This is all to say that when checking for counterfactual dependence between an event and an omission, we investigate whether the effect would occur had the event which is the omission actually occurred. Applied here, we'd ask whether Megan's death would have occurred, if Susan *had* turned the steering wheel. Given that it's true that had Susan turned the steering wheel Megan would not have died, there is counterfactual dependence between Susan's failure to turn the steering wheel and Megan's death. As a result, the counterfactual analysis attributes omissive causal responsibility to Susan.

However, we soon notice that the counterfactual analysis attributes omissive causal responsibility to people other than Susan. To see this let's add another person to the story. Suppose that Susan has a passenger, Ali, in the backseat of her car. Ali is sat directly behind Susan; his view of the street is obstructed and he cannot see Megan crossing the street. Even though Ali omitted in exactly the same way as Susan — they both failed to turn the steering wheel — presumably Ali is not causally responsible for Megan's death. Indeed, it seems like

Ali's omission doesn't make him causally responsible for anything at all. However, if our attributions of omissive causal responsibility are identical to causal attributions given by the counterfactual analysis, then Ali is omissive causally responsible. There is counterfactual dependence between Ali's omission and Megan's death, for had Ali not failed to move the steering wheel, Megan would not have died.

In fact, basing our attributions of omissive causal responsibility on the counterfactual analysis would mean that *any* agent who fails to move the steering wheel would be omissive causally responsible for Megan's death since the closest worlds in which someone moves the steering wheel are worlds where Megan survives. Wonder Woman's failure to move the steering wheel makes her omissive causally responsible in virtue of the fact that had she not failed to perform that action, Megan would not have died.

The problem is compounded once we realise that it's not only failures to turn the car's steering wheel that can endow an agent with omissive causal responsibility, other seemingly irrelevant omissions can too. For instance, Susan's failure to stop for coffee at the start of her drive that day also makes her omissive causally responsible for Megan's death, since had she stopped for coffee Megan would have safely crossed the road some time before Susan drove by in her car. Manifestly then, the counterfactual analysis grants far more instances of omissive causal responsibility than intuition would allow.

Menzies (2004) recognised this problem with the counterfactual analysis in the context of metaphysical causation. He argued that the counterfactual account is too profligate in generating causes for any given effect, the worry has since been referred to as the problem of profligate causes. Menzies characterises the problem as a distinctly metaphysical concern, but as the above discussion demonstrates, it also generates concerns in the moral domain, for if omissive causal responsibility is assigned on the basis of counterfactual dependence, then there would be far more instances of omissive causal responsibility than intuition would allow for.[39]

[39] In Chapter 4 I said I was leaving interventionism behind, one of the reasons why I chose to do this is because it fails to provide an account of omissive causal responsibility for the same reasons as the simple counterfactual account. For simplicity's sake, I therefore chose to focus on the simple counterfactual analysis. Let me briefly outline why interventionism fails. Interventionism will also produce a near infinite number of omissive causal relations if we put these omissions into a causal model. Once an omission is represented in a model, intervening on the variable representing the omission by switching it from 'off' to 'on' will most of the time produce an associated change in another variable in the model. Hence, the model will say that there's a causal relationship between the omission and the other variable. If we manipulated the variable representing Ali's failure to turn the wheel from off to on, then it would bring about a change in the variable representing Megan's death.

Now, one might respond to this worry by simply biting the bullet. Specifically, one might be tempted to accept the account's overly permissive attributions on the grounds that such permissiveness is not problematic for the purposes of assigning moral responsibility. In the previous Section, I noted that those who accept CR and MRO need to establish omissive causal responsibility *for the purpose* of holding agent's morally liable for outcomes. The response might go that the counterfactual analysis allows us to do precisely this. Given that the analysis grants a near infinite number of omissive causal responsibility attributions, we can hold agents morally liable, and insofar as the view fulfils its purpose in this regard, we're justified in accepting far more omissive causal responsibility than we would intuitively think acceptable.

However, I don't think we should bite the bullet here. This is because it seems to me that the account's permissiveness undermines the usefulness of the concept of omissive causal responsibility. When it comes to moral responsibility attributions, omissive causal responsibility is of practical and theoretical interests to us because it picks out those omissions, from an infinite number of other nondoings, that are especially attributable to agents as genuine failures to act. By assigning omissive causal responsibility to an agent we imagine that the omission in question is attached or *authored* by her in such a way that it makes sense to take it as a legitimate basis for some kind of moral judgement about her. In this sense, theories of omissive causal responsibility should identify those omissions which are reflective of a person's agency.[40] The kind of agency that omissive causal responsibility ought to track, I suggest, can be put in terms of *answerability*.

5.7 Fixing the Target: Answerability and Omissive Causal Responsibility

In this Section I will outline the concept of answerability and its connection to omissions, but first let me say what role answerability is supposed to be playing in my discussion. I'm not

---

Of course, we could try to restrict the kind of omissions that should be represented in the model, but this requires going further than the original account. It would also require providing arguments in favour of certain restrictions over others.

[40] In the previous Section, I noted that my analysis of omissive causal responsibility will include both intentional and unintentional omissions. It's no mystery as to how intentional omissions have the ability to reflect one's agency; forming the intention not to act and manifesting that intention is straightforwardly an expression of one's judgemental self. But it might be less clear how agency can be tied to unintentional omissions. To this I say that some unintentional omissions can manifest agency, because sometimes people omit as a result of failing to respond to reasons they ought to have responded to, and in such cases, their failure to act reveals something about the type of person they are. When Kate forgets to call ger grandmother to wish her a happy birthday, she unintentionally omits because she failed to respond to reasons she ought to have responded to, such as, not calling her grandmother will hurt her grandmother's feelings. By unintentionally omitting, Kate expresses something about the type of person she is, albeit in a small way.

supposing that answerability is a *theory* of omissive causal responsibility. That is, I don't suppose that the concept can provide necessary or sufficient conditions for omissive causal responsibility. This is primarily because theories of omissive causal responsibility aim to provide conditions which establish a *causal* link between omissions and outcomes, and the concept of answerability is not in the game of supplying such a link. Instead, I am arguing that answerability plays a more fundamental role: answerability *fixes the target* for theories of omissive causal responsibility. It tells us what kind of omissions a theory ought to be picking out as able to ground attributions of moral responsibility. Identifying this target provides a success condition for theories of omissive causal responsibility; it identifies what a theory ought to be getting at.

With the role of answerability clarified, let me explicate the concept. Angela Smith (2017) says that an agent is answerable when we are entitled to request a justification from her about why she behaved the way she did. And what makes such a request for justification appropriate is the fact that the agent's behaviour is genuinely authored by her or reflects her own judgement in a way that expresses her agency. In terms of omissions, one would be answerable for one's failure if someone were justified in asking why one failed to act. To illustrate, suppose that Disha breaks a lifetime habit by omitting to make porridge for breakfast because that morning she fancied toast. Disha is answerable for her omission insomuch as her partner, for example, could intelligibly ask Disha why she's not making porridge. There's a sense in which Disha's partner would be entitled to request information about why she refrained. Importantly though, Disha is not answerable for all of her omissions. She's not answerable to anyone when, sat at her desk in the UK, she fails to stop a mugging happening on the New York Subway. There's no sense in which anyone would be entitled to ask Disha why she failed to stop the mugging. Such a request for information would be unintelligible because it does not make sense to ask an agent to justify a failure that bears no relation to her own agency.

As the example illustrates, answerability is not a moral evaluation. The mere fact that one is answerable for an omission does not yet show that one is morally responsible for the effects of that omission. Disha is answerable for her failure to make porridge, yet she is not morally responsible for anything. Disha's omission has morally neutral consequences, thus there's nothing we can hold her morally responsible *for*. Moreover, even if the omission for which the agent is answerable does produce morally significant consequences, she still may not warrant praise or blame, for surely this depends upon how well or poorly she answers the justificatory request. In other words, a moral evaluation (as opposed to a causal evaluation) will depend

upon the quality of the justificatory reasons she is able to offer in explanation for her omission — an inadequate response will likely make her blameworthy, while a satisfactory answer could let her off the hook.[41]

With all that said, answerability is importantly linked to moral responsibility. Although being answerable does not by itself show that an agent warrants any substantive moral judgement, it does serve as the basis for such judgements. Smith (2017) describes answerability as a "gateway notion": it allows us to pick out those omissions for which the agent is responsible in the most basic sense, and in turn those omissions may serve as the basis for some kind of moral response if what the agent does or intends is morally significant (p. 51).

If this is the right way to fix the target for theories of omissive causal responsibility, then the counterfactual account's permissiveness is a problem after all. In granting a near infinite number of omissive causal responsibility attributions, the analysis allocates omissive causal responsibility to those failures for which we are answerable *and* those for which we are not answerable. It therefore does not track the target phenomenon.

5.7.1 Answerability and People on the Moral Margins

Before moving on, I want to take a moment to respond to a potential objection to understanding omissive causal responsibility as related to answerability. Put simply, the objection says that answerability is too narrow of a target to capture all cases of omissive causal responsibility. I've said that someone is answerable when we're entitled to ask for justification about why they failed to act. Occasionally however, we might think that asking someone to justify themselves would be inappropriate, yet nevertheless, they strike us as being causally responsible. The worry is borne out in cases involving small children and people with severe mental impairments.

To illustrate, let me focus on a case involving small children. Imagine that in order to attend a nursery trip pupils must get written permission from their parents. A teacher puts consent forms into the bags of all her 3-year-old pupils, and instructs them to pass the consent forms onto their parents. One pupil, Eva, forgets to give her parents the consent form. On the day of the school trip, Eva is not allowed to accompany her classmates. Some might have the intuition

[41] There is another sense in which answerability is not a moral evaluation. One can be answerable outside of the moral domain such as the epistemic domain or the aesthetic domain. For instance, one would be entitled to ask someone why they hadn't formed the belief that 2+2=4 (supposing it was reasonable for us to expect them to have formed that belief).

that Eva's failure to hand over the consent form makes her causally responsible for the fact she's not allowed on the school trip.[42] Yet, it's not obvious that Eva is answerable for her omission. It seems like neither her parents nor her teacher would be entitled to ask Eva to justify why she didn't pass on the consent form. Demanding justification seems like an unreasonable thing to do given that 3-year-olds lack the capacities which would make remembering to pass on a consent form an easy task. If it is true that Eva is not answerable, then according to my view, she cannot be held causally responsible for the consequences of failing to pass the consent form on to her parents.

There are two kinds of responses one could make here. The first kind of response explains away these cases as *problem* cases. It begins by pointing out that I am seeking a notion of omissive causal responsibility that specifically grounds moral responsibility evaluations. By assigning omissive causal responsibility to an agent, I have suggested that we signal that the omission in question expresses her agency in such a way that it can form a legitimate basis for moral assessments about her. It's widely recognised that small children (and people with severe impairments) do not have the required capacities (whatever they may be) to be the basis for such moral assessments. Indeed, theories of moral responsibility are often judged on whether they can exclude and explain why children's actions and inactions are not appropriate objects for moral judgement. In light of the fact that small children cannot be held morally responsible, it's not a problem that my understanding of causal responsibility entails that children cannot be held omissive causally responsible.

The second response doesn't explain away the cases, instead the response holds onto the claim that small children can be omissive causally responsible by pushing back against the thought that children cannot be answerable for their omissions. In some instances, children, even as young as 3, can be answerable for their behaviour. When a child behaves poorly by, for example, drawing on the walls with permanent marker, we can easily imagine a parent asking 'Why did you do that?!'. There's nothing inappropriate or unreasonable about this reaction (indeed it might be more bizarre if the parent did not ask for a justification). The reason such a reaction is not unreasonable is because the child is connected to the drawings on the wall in a

[42] My intuition about this case is not clear. Specifically, I'm uncertain as to whether Eva is omissive causally responsible for her not being allowed on the school trip. Perhaps it would be more intuitive to say that the teacher is causally responsible, or that the parents' failure to check Eva's school bag makes them causally responsible. Then again, perhaps these latter intuitions are tracking judgements about who's to blame rather than who's a cause.

way that makes these answerability demands intelligible. Her action is authored by her — it reflects an emerging sense of agency.

We can apply this sort of thinking to omissive cases like Eva's. Perhaps it doesn't seem unreasonable after all to ask Eva why she didn't hand over the consent form to her parents. It certainly seems like an intelligible demand when contrasted with other demands; say, demands for Eva to justify her eye colour, her sense of humour, her heartrate. These demands are obviously unintelligible. It would make no sense to request that she justify these features because they bear no relation to her agency. Whereas forgetting to hand over the consent forms could reflect something about Eva — her lack of desire or enthusiasm to attend the nursery trip, for example. If this is true — in particular, if it's true that small children can be answerable for their failures to act — then they can also be omissive causally responsible.[43]

For what it's worth, out of these two responses to the objection, I favour the latter response. A full-blown analysis of answerability and in particular the conditions under which it would be appropriate to demand justification would need to be provided in order to thoroughly make the response. This is sadly outwith the scope of the Chapter. Nevertheless, the discussion above suggests to me that there's at least a minimal sense in which small children can be answerable for that they do. To borrow David Shoemaker's (2015) words, small children seem to possess a "burgeoning answerability" (p. 76) that is sufficient to ground claims about omissive causal responsibility.

5.8 The Success Conditions for a Theory of Causal Responsibility

Here is a good place to pause to clarify the dialectic. In the preceding discussion, I have suggested that the appropriate target for theories of omissive causal responsibility are omissions which reflect a person's agency. For it is only these agency-indicating omissions that have the ability to ground assessments of moral responsibility. I've cashed out agency in terms of answerability. This discussion generates some success conditions for a theory of omissive causal responsibility that have hitherto been absent from the philosophical literature. A successful theory will be one that is able to pick out those failures for which we are

[43] It's important to bear in mind, however, that this does not entail that Eva would be blameworthy for her omission. Establishing omissive causal responsibility does not entail that we have established moral responsibility. Causal responsibility is a necessary but not sufficient condition on moral responsibility; Eva might fail to meet these other conditions, meaning she can't be held morally responsible.

answerable, and to establish a causal link between those failures and the outcomes for which we want to attribute moral responsibility.

Now given that this is what we want from our accounts of omissive causal responsibility, it's no surprise that the simple counterfactual analysis cannot supply one. After all, the counterfactual analysis is a theory of causation and most theories of causation, have tended to look for the conditions that are, as it were, minimally sufficient for a relationship to qualify as causal. That is, the focus has been largely on finding the conditions that distinguish causal from noncausal relationships. Given this orientation, the simple counterfactual analysis is not equipped with the resources to make further discriminations across the causal/noncausal line, such as, whether certain causal omissions also express a person's agency.

Still, there will be occasions when the counterfactual test gets attributions of omissive causal responsibility right. This happens when an omission qualifies as a cause and also happens to be an omission for which an agent is answerable. We've seen such a coincidence in the account's treatment of **Driving**. According to the counterfactual test, Susan's failure to turn the steering wheel makes her causally responsible for Megan's death in virtue of the fact that the effect counterfactually depended on her omission. Now, as it happens, Susan's omission is the type of nondoing for which she is answerable; it would be quite appropriate for us to ask Susan to justify why she did not move the steering wheel. The counterfactual test has, albeit somewhat accidentally, picked out an omission that reflects a person's agency, and as a consequence, it gets the attribution of omissive causal responsibility right. The problem is that for most of the omissions the test identifies as causal are not the sorts of omissions for which we are answerable. In these sorts of cases the counterfactual test misfires as a test for omissive causal responsibility. This is illustrated when we looked Ali, the backseat passenger, and Wonder Woman. Although Megan's death counterfactually depends on both of their omissions, they don't happen to be the sorts of failures that require a justification. For this reason, we don't see Ali or Wonder Woman as bearing causal responsibility.

Now that we have a clearer idea of the target and success conditions for theories of omissive causal responsibility, we are in a better position to seek out the conditions for it. With this in mind, I will now move to explore a second causal theory as a contender for omissive causal responsibility. I call this causal theory the 'normative analysis'. I explain my motive for focusing on the normative analysis in the next Section.

5.9 The Normative Analysis

In her influential paper ‘Causation by Omission: A Dilemma’ (2005) Sarah McGrath sets out to discover the basis on which common sense singles out some omissions as causes but not others. For example, why does it appear to be that my failure to feed my own fish is a cause of the fish’s death, but Bono’s omission to feed my fish is not a cause of their death? After evaluating and rejecting several proposals, McGrath defends the view that such a basis for discrimination is formed of normative considerations — “my proposal involves embracing the view that causation by omission has a normative component” (p. 125). More specifically, she argues that the type of component needed in a theory of causation by omission is a notion of *normality*.

I have chosen to focus on the normative analysis for two chief reasons. Firstly, it comes much closer than the counterfactual analysis to capturing the target for omissive causal responsibility theories. As we’ll see, the notion of normality initially appears to play a crucial role in distinguishing the types of omissions for which we are answerable from those which we are not. Secondly, as I’ll shortly show, many philosophers have appealed to normative considerations to identify those failures that are relevant from the point of view of ethics. It therefore seems fair to give a causal account that incorporates norms a good airing. Despite its initial promise however, I’ll ultimately argue that the normative analysis does not always identify the omissions for which we are answerable, and therefore, does not capture the target phenomena. Specifically, it does not ascribe omissive causal responsibility to agents who we think are answerable for their failures, making the view too restrictive. Furthermore, I’ll present a wider worry about the view which concerns the general strategy of appealing to normative considerations to determine omissive causal responsibility. Simply put, my complaint is that norm violations do not track answerability. I show this with a case involving pernicious norms.

In what follows, I’ll begin by laying out the normative analysis. Next, I’ll put the view to work as a theory of omissive causal responsibility, and I’ll demonstrate that the view gets a lot of cases right that the counterfactual analysis gets wrong. Then I’ll consider some difficulties faced by the account, before drawing a general conclusion about the use of normative considerations in attributing causal responsibility.

5.9.1 Outlining the Normative Analysis

According to McGrath, an omission is a cause only if it's occurrence would have normally prevented the effect from occurring. Conversely, omissions that, had they occurred, would not normally prevent the effect are not causes. McGrath states the view as follows:

> An omission, *O*, is a cause of some effect, *E*, iff both *O* and *E* occur, and:
>
> the kind of event of which *O* is the omission is a *normal would-be preventer* of *E* (if e were to be prevented, it would have been normal for an event of this type to prevent it). (2005, p. 134) [44]

The account takes the counterfactual test as its foundation. To check for causality, one needs to consider whether the effect would occur had the event which was the omission occurred. The difference here is that counterfactual dependency is not sufficient for causation; the omission must also be the type of event which, had it occurred, would normally prevent the effect from occurring. The addition of this normal quantifier makes the normative analysis far more restrictive in its identification of causal relations involving omissions. Unlike, the simple counterfactual analysis, the view does not recognise a near infinite number of omissions as causes, on the contrary, only those omissions which are normal would-be preventers count as causes.

Before seeing the view in action as an account of omissive causal responsibility, there are a couple of important features to note about the normative analysis. Firstly, the concept of normal being employed by McGrath is intended to be wide-ranging as to include both statistical and prescriptive norms. I introduced the kind of norms McGrath is incorporating in Chapter 2, to recap: to say something is normal in the statistical sense is to say that it conforms to a statistical standard. For example, in Glasgow the winter months are generally rainy and overcast, so if Glasgow were to have a sunny, dry winter, the city's weather would violate a statistical norm. By contrast, to say something is normal in a prescriptive sense, is to say that thing follows a prescriptive rule, and these rules are constituted by the way things *ought* to be or are *supposed* to be. Prescriptive norms constitute a huge heterogenous category. Some norms are moral; for example, it's generally believed that people are supposed to keep their promises, even if there are no explicit laws or rules demanding this behaviour. There are also etiquette norms, contractual norms, legal norms, functional norms, epistemic norms, cultural norms and so on,

[44] McGrath is supplying an account of omissive causation generally, she's not concerned with grounding moral responsibility attributions in particular.

all of which generate expectations about how things are supposed to be. McGrath appeals to the full range of norms in her analysis.

Secondly, McGrath recognises that there may be conflicting standards of normal such that an omission is a normal would-be preventer according to one standard, but not a normal would-be preventer according to another standard. To illustrate, suppose I promise to collect a friend from the airport, but being an absent-minded person who regularly forgets her obligations, I arrive at the airport two hours late to be met by an annoyed friend. My failure to arrive to the airport on time is a normal would-be preventer of my friend's annoyance according to the standards set by promise-making and norms of friendship. However, my failure is not a normal would-be preventer according to the standards set by my own forgetfulness — it would be statistically unlikely for me to collect my friend at the expected time. Oddly then, my omission is a cause according to some standards, but not a cause according to another standard. McGrath anticipates this implication: she states that when such conflicts occur the omission is a cause simpliciter if it counts as a would-be prevented according to at least one of these standards. Hence, my failure to collect my friend on time is a cause simpliciter since it violates at least one standard of normality; namely, norms of promising and friendship.

5.9.2 The Normative Analysis and Omissive Causal Responsibility

Having outlined the key features of the view, let's now consider how it fairs as an account of omissive causal responsibility. Let's look at **Driving** again:

> Susan is driving in a suburban neighbourhood. Her hands are on the steering wheel when she observes a pedestrian (Megan) crossing the street approximately 100 feet in front of her. She knows she could easily move the wheel to avoid hitting Megan without veering outside of her lane. Instead, she deliberately remains motionless. Her car hits and kills Megan.

What would the normative analysis say about such a case? According to the account, Susan's failure to move the steering wheel makes her omissive causally responsible for Megan's death in virtue of the fact that her omission is the type of event that, had it occurred, would have normally prevented the effect. In particular it would be morally and legally normal for drivers to prevent collisions by manipulating the steering wheel, guiding the car, breaking, etc., especially when doing so poses no risk to their own safety. Furthermore, not everyone's failure to move the steering wheel makes them omissive causally responsible for the death, because not every failure is a normal would-be preventer of the effect. For example, Ali — the backseat

passenger — is not omissive causally responsible according to the normative analysis because their turning the wheel would not normally prevent the effect. It would be statistically abnormal for backseat passengers, who are unaware of a pedestrian's presence, to leap across the car and turn the steering wheel. And there's no sense in which one ought to act like this either.

As the example illustrates, norms seem to play a crucial role in an analysis of omissions. In fact, a normative approach to omissions has been around long before McGrath popularised it in the causation literature. Many authors have invoked normative considerations to draw a line between the infinite number of nondoings occurring at every moment and those omissions which seem relevant from the point of view of ethics. For example, in *Responsibility and Fault* (1999) Tony Honoré argues that an omission is a nondoing, but not every nondoing is an omission. According to Honoré, "to omit implies one ought to have done what was not done […] an omission violates a norm" (p. 47). Furthermore, in a number of thought-provoking articles, Patricia Smith (1990) aims to distinguish between conscious and unconscious omissions. Smith argues that difference lies in the fact that the very concept of an unconscious omissions depends upon what she calls a "contextual standard of reasonable expectation" (p. 163). On her view, in order to determine whether an agent A has unconsciously omitted to do X, all we need to know is whether the agent did not think of doing X at the time, and whether it was reasonable to expect A to do X in the context. Unconscious omissions, then are simply those unconscious nondoings that violate or deviate from what we expect the agent to do. Although Smith does not employ the language of normality, it's clear that some such notion is driving what she thinks these expectations consists in.

More recently, Randolph Clarke (2014) has said that when it comes to separating omissions from other ordinary nondoings, the difference between the instances that count as omissions and those that don't, does not lie in what kind of entity exists in the two cases, but in whether an action of the sort that is absent was demanded by some norm. He says that "one seldom if ever counts as having omitted to do something unless there was some norm, standard, or ideal that called for one's so acting" (p. 33).

So the idea that omissions and norms are tightly connected is not new. Still, I've chosen to focus on McGrath's delineation of the normative approach because her account is after all a theory of causation. It therefore translates well as an account of omissive *causal* responsibility. Moreover, unlike some other accounts, McGrath spends time spelling out exactly what kinds of normative considerations she considers to be relevant. Supplying this detail is a crucial step

if we're to put the view into practice to check for the existence of causal responsibility. With that said, not much hangs on McGrath's formulation *per se*. The difficulties I raise about McGrath's normative analysis are not generated by any features particular to only her account. The difficulties I raise will be faced by any approach which appeals to norms in order to establish omissive causal responsibility. It's these difficulties to which I now turn.

5.9.3 Problems with the Normative Analysis

The first problem with the normative analysis concerns its failure to attribute causal responsibility when agents are answerable. This fault is drawn out when we attend to cases involving commendable omissions (as opposed to blameworthy omissions).[45] To illustrate the problem, consider the following example:

> **President:** In the midst of a war, the President is strongly encouraged by her special advisors to authorise a nuclear attack on country X. The President rejects their advice, and decides not to authorise the nuclear attack, as a consequence, the 50 million residents of country X remain unharmed.

It's natural to think that the President's omission makes her answerable, for it would be completely appropriate to ask her why she didn't authorise the attack. In light of this, we ought to be able to say that her failure to launch the attack makes her causally responsible for the residents of X remaining unharmed. The normative analysis, however, cannot establish this causal link. This is because the President's omission is not the type of event that had it occurred would normally prevent the effect from occurring. A nuclear attack is not the type of thing that would normally prevent 50 million residents of a country from continuing to remain unharmed. It would be extremely unusual in a statistical sense for a country to be destroyed as a result of a nuclear attack, and it's hard to imagine a sense in which authorising a nuclear attack would adhere to the way things ought to be or are supposed to be. As a result, when the President omits, she is not causally responsible for anything according to the normative analysis. And the problem isn't exclusive to **President**; it applies to any cases where one fails to do something in circumstances where the situation does not call for it to be done. Thus, the normative analysis does not establish causal responsibility when agents are answerable for their omissions, making the view too restrictive.

---

[45] The objection that the normative analysis cannot handle cases involving commendable omissions has been argued for before by Carolina Sartorio (2007, p. 756). Though Sartorio does not put the objection in terms of answerability.

The second worry I have about the normative analysis is much wider. It concerns the general strategy of appealing to normative considerations to determine omissive causal responsibility. Simply put, my complaint is that norm violations do not track answerability. The kind of cases that clearly bring out my complaint are those involving violations of pernicious norms. To see the worry, let us suppose that there exists a patriarchal norm according to which it is normal for women (and not men) to carry out domestic labour. Suppose that Tamsin and Bill live together and suppose that they're throwing a dinner party. On entering the house, one dinner guest loudly remarks on how dusty their dining room is inducing a pang of shame in Tamsin and Bill.

Now, the fact that there's a gender norm according to which it would be normal for Tamsin to undertake the domestic labour, means that her omission to dust is a normal would-be preventer of Tamsin and Bill's shame. According to the normative analysis then, Tamsin's failure to do the domestic labour makes her omissive causally responsible for the fact that the pair feel ashamed. Furthermore, Bill's failure to dust is not a normal would-be preventer, so he is *not* causally responsible for their shame.[46]

What should we say about this assignment of causal responsibility? Well, it will depend upon whether we think Tamsin and Bill are answerable for their failures. If we think that Tamsin is answerable, and Bill is not answerable, then the normative analysis has got the assignment correct. The problem is that this just doesn't seem to be the case. Depending on the details of the story, there are two plausible options in terms of who is answerable. Suppose that the dinner party is a formal occasion and that the layers of dust are especially noticeable. Then there might be grounds for requesting why the house hadn't been dusted before the guests' arrival. But notice that if this is true, it would appropriate to request this information from *both* Tamsin and Bill. For surely Bill would be answerable for the same reasons Tamsin is answerable, even with the existence of the patriarchal norm.[47] On this reading of the case, we should hold both agents

[46] I can't see a sense in which Bill's failure to dust violates any other norm. Given the patriarchal norm, there's no social or cultural standard according to which it would be normal for him to dust. Nor does he violate other moral, prudential, epistemic, etc, standards by not dusting his dining room.

[47] One might want to push back against this claim. Specifically, one might say that *given* there is a patriarchal norm according to which Bill is not expected to undertake the domestic labour, it would be inappropriate to ask Bill why he hadn't dusted the dining room. But this response would turn answerability into a relativised concept. It would mean that whether one is answerable depends upon the norms that happen to be in operation at a particular time and place. This is not how we should conceive of answerability. An agent can be answerable even if their conduct doesn't violate any of the norms that happen to be in operation. For instance, imagine I volunteer some of my time to work at a

causally responsible for the shame they feel, hence the normative analysis gets the wrong verdicts.

On a different version of the story, it would seem like neither Tamsin nor Bill are answerable. Suppose that the dinner party was an informal affair with close friends, and that the dust is hardly noticeable. Then it seems that Tamsin and Bill have nothing to answer for; it would be quite inappropriate for one of their dinner guests to demand justification for why the dining room hadn't been dusted. According to this version of the story, it would be wrong to hold Tamsin (and Bill) causally responsible for the effects of their failure to dust. So whatever side we decide to come down on in terms of establishing answerability, the normative analysis fails to reflect these judgements in its assignments of causal responsibility. Much like the simple counterfactual analysis, the account goes wrong in its allocations of causal responsibility by way of failing to track the target phenomena.

5.10 A New Account to Omissive Causal Responsibility

Let's take stock about what's been said thus far. I began by laying out what I take to be a challenge for those who accept CR and MRO; namely, to make sense of the idea that agents can be causally responsible for outcomes when they fail to act. In Section 7 and 9, I looked at what would happen if we based our attributions of omissive causal responsibility on the simple counterfactual analysis and the normative analysis. I argued that both views fail as theories of omissive causal responsibility because they do not pick out the target phenomena — omissions for which we are answerable. In this Section, I will present an alternative account of omissive causal responsibility. In short, my account says that an agent is omissive causally responsible for an outcome if and only if the occurrence of the outcome counterfactually depended on the occurrence of the omission, *and* the relationship between the outcome and the omission is a causally stable one. More formally:

> **Omissive Causal Responsibility:** an agent is causally responsible for some outcome when they fail to act, if (i) the outcome counterfactually depends on the omission and (ii) the relationship between the omission and the outcome is causally stable.[48]

---

charity bookshop. Someone would be entitled to ask me why I engage in this practice, even though my volunteering does not violate any norms.

[48] There might be a way to spell out this view without condition (a). The condition could be subsumed under condition (b) if spelled out in sufficient detail. However, for clarity and presentational purposes, I present the two as distinct, separate conditions.

Condition (i) is the simple counterfactual test. However, as I noted in Section 7, on its own (i) would produce a near infinite number of omissive causal responsibility attributions which generates a problematically permissive account. Adding condition (ii) avoids an overly inclusive view by restricting those omissions which make one causally responsible to those that express stable causal relationships. Importantly, condition (ii) is not an *ad hoc* addition whose purpose is to merely limit the amount of omissive causal responsibility out there in the world. Rather this condition restricts omissive causal responsibility to the right kind of omissions — ones for which we are answerable. I will say more about the connection between answerability and stability later, for now it's worth pointing out that I do not see stability and answerability as extensionally equivalent. The connection between the two concepts is much less strict, I see stability as a way of tracking answerability. In this sense causal stability is being deployed as a kind of epistemic tool to identify the target for theories of omissive causal responsibility — agency-indication omissions.

With the formal conditions outlined, I will now turn to delineate the view in more detail, starting with a discussion about causal stability.

5.10.1 Causal Stability

First introduced by David Lewis (1987) under the name "sensitivity" (p. 184), stability has been a relatively overlooked dimension of causation, but contemporary discussions by philosophers and psychologists have begun to popularise the concept.[49] Broadly, a causal relationship is stable to the extent that it holds across a variety of background circumstances. If the causal relationship holds in a wide variety of background conditions, then it is relatively stable, conversely, if it fails to hold in a variety of background conditions, then it is relatively unstable.

I introduced stability in Chapter 3. There I said that stability was adopted by interventionists as a criterion to determine the aptness of a model, apt models being those that express the appropriate number of variables to secure stable causal relationships. The kind of stability cited in Chapter 3 is broadly similar to the notion I am using here insomuch as both concern what will happen to causal relationships in different circumstances. But the specifics are quite different. In Chapter 3, stability applied to model construction, but since I've left interventionism behind, I'm no longer interested in how stable causal relationships are *relative*

[49] For example: Laura Franklin-Hall (2016), and Nadya Vasilyeva, Thomas Blanchard and Tania Lombrozoa (2018)

to a model, but rather how stable causal relationships are more generally. Furthermore, in Chapter 3, I noted that interventionists understood stability to be a virtue of a model because stable relationships offer up information about how we can control the world around us. Given this was the motivation for appealing to stability, I claimed that stability needed to be measured by enacting specific relevant kinds of manipulations on background conditions implicit within a model. I then suggested that what makes a manipulation relevant for measuring stability depended upon whether these manipulations were 'normal manipulations'. Normal manipulations being those that either keep the background conditions normal or changed the abnormal background conditions to normal conditions. But in this Chapter, my primary aim is not to uncover causal information pertinent for manipulating and controlling causal relationships. As such, there's no need for me to be so discerning about the kind of changes we need to attend to in order to test for stability. This is to say that there's no grounds for me to attend to only normal manipulations. I'm using the notion far more generally in a sense which is not characterised by its ability to yield causal information useful for manipulation. My measurement of stability will therefore be different than the one used in Chapter 3.

Stability was characterised in broad brush strokes in Chapter 3, here I expand more thoroughly on it, paying special attention to how one measures the concept. As before, I take the notion of stability from James Woodward (2006), although my detailed understanding of it and the use to which I will put it is somewhat different from his.[50] For Woodward, when assessing for stability the guiding idea is that we are to consider possible changes from the actual world in which the cause event occurs and then ask whether the effect event also occurs in these possible worlds. In other words, we imagine that the cause event occurs, wiggle certain background conditions, and then look to see whether the effect would also occur. For Woodward, this test entails asking how many worlds the following counterfactual would hold in:

$O(c) \mathbin{\square\!\!\rightarrow} O(e)$

In English: if *c* were to occur then *e* would occur. To the extent that the counterfactual holds in many worlds the causal connection is stable and to the extent it fails to hold in many worlds

[50] For one thing Woodward rejects Lewis's possible world semantics, but I assume such an approach to determining the truth value of counterfactuals. Furthermore, Woodward's check for stability makes use of both of Lewis's counterfactuals for causation. That is, he makes use of both the $O(c) \mathbin{\square\!\!\rightarrow} O(e)$ and $\neg O(c) \mathbin{\square\!\!\rightarrow} \neg O(e)$ counterfactuals. But I only focus on the $O(c) \mathbin{\square\!\!\rightarrow} O(e)$ counterfactual because doing so is sufficient for establishing stability in the context of omissive causal responsibility.

the causal connection is unstable. In the next Section, I'll say something about what specific kinds of worlds are relevant for assessing stability. For now, let us follow the conventional constraint by restricting the relevant types of worlds to those that are close-by or not too dissimilar from the perspective of the actual world.

Notice that the counterfactual being invoked by Woodward is identical to the one Lewis uses in his 1973 theory of causation. I noted that this counterfactual is often overlooked because *O(c)* $\Box\!\!\rightarrow$ *O(e)* will automatically be true so long as *c* and *e* occur, thus, in such cases, whether *e* is counterfactually dependent upon *c* will be determined by the truth value of Lewis's other counterfactual ¬*O(c)* $\Box\!\!\rightarrow$ ¬*O(e)*. Here Woodward is suggesting that the often overlooked counterfactual can actually yield interesting information regarding certain properties of causal relations; namely, their stability. But crucially, uncovering this information requires that we do not merely check to see whether the counterfactual is true (as Lewis asks us to do) but also to check in what range of worlds the counterfactual is true. Woodward puts the idea as follows: "while the counterfactual [*O(c)* $\Box\!\!\rightarrow$ *O(e)*] is trivially true if *c* and *e* occur, it is a further nontrivial question whether and to what extent [the counterfactual *O(c)* $\Box\!\!\rightarrow$ *O(e)*] would continue to hold under various departures from the actual circumstances" (p. 4).[51]

To illustrate how one measures causal stability, compare the following claims:

> *(1) Suzy throwing the rock at the window caused the window to shatter*
> *(2) The big bang caused the window to shatter*

To assess the stability of the causal relationships expressed by these claims we should translate the statements into the associated counterfactual of the form *O(c)* $\Box\!\!\rightarrow$ *O(e)*. This amounts to placing the occurrence of the cause in the antecedent and the occurrence of the effect in the consequent:

> *(1a) If Suzy were to throw the rock, then it would be the case that the window would shatter*

[51] It's worth noting that although stability makes use of Lewis's framework for causation, the check for stability does not equate to a check for causation. Causation would require considering the negated counterfactual as well.

(2a) If the big bang were to occur, then it would be the case that the window would shatter

Once the associated counterfactuals have been identified, we then move to consider in how many worlds these counterfactuals will be true. Looking at (1a) first, we can see that there are a broad range of nearby worlds in which the antecedent and the consequent are both true. The counterfactual would hold in worlds that contain relatively trivial changes to the background conditions; for example, had Suzy been wearing different clothes, had she had brown eyes instead of blue, or had she thrown with her left hand rather than her right, the window would still have shattered. More interestingly, the counterfactual would also hold in worlds that represent larger departures from reality; had Suzy thrown the rock a year earlier than she actually did or had she thrown it from five feet to the left of her actual position, the window would still have shattered. There will be certain changes in the background conditions that can render the counterfactual false. Had the windowpane been made from a special type of reinforced glass then the window might not have shattered even though Suzy threw a rock at it. Still, it's hard to imagine many scenarios where (1a) would be false since in most contexts throwing a rock at a window will break the window.

In comparison to (1a), there are many departures from the actual circumstances that would make (2a) false. That is, there are many scenarios in which the big bang occurs and the window does not shatter. These scenarios include those that would make (1a) false, like the windowpane being made from reinforced glass, but they would also include any departure from actuality that occurred after the big bang that resulted in, say, the non-existence of Suzy or the non-existence of that particular window. And given the enormous range and diversity of possible changes that would result in these scenarios we get the impression that (2a) would fail to hold in many worlds. In light of the fact that (1a) holds in many more worlds than (2a), the causal relationship between Suzy's rock throwing and the shattered window is significantly more stable than the causal relationship between the big bang and the shattered window.

That's how the concept of stability can be applied to 'positive', bone fide events. Let me now turn to demonstrate how it can be applied to causal relationships involving omissions. Consider the following claim:

(3) Not being abducted by aliens caused me to drink this very coffee.

Let's step back a second to consider what Lewis's treatment of causation would say about this claim. According to the simple counterfactual analysis, (3) is true if the following counterfactuals are true:

> *(3a) If I were not abducted by aliens, then I would have drunk this very coffee*
> *(3b) If I were abducted by aliens, then I would not have drunk this very coffee*

(3a) expresses the counterfactual $O(c) \Box\!\!\rightarrow O(e)$. We got to (3a) by placing the occurrence of the omission in the antecedent, and the occurrence of the effect in the consequent. (3b) expresses the negated counterfactual $\neg O(c) \Box\!\!\rightarrow \neg O(e)$. We got to (3b) by putting the non-occurrence of the omission in the antecedent (which equates to its occurrence), and the non-occurrence of the effect in the consequent. When testing for causation we typically focus on assessing the truth value of (3b). In this instance (3b) would presumably be true because the closest worlds in which I am abducted by aliens are also worlds where I don't drink the coffee. In testing for stability, I'm asking us to shift our attention to (3a). But importantly when attending to (3a) we are not merely checking whether it is true (as Lewis's check for causation demands) but also in what range of worlds (3a) is true. The larger the range of worlds the more stable the causal connection, whilst the smaller the range of worlds the less stable the causal connection.

So as with claims involving ordinary positive events, when the causal relationships have omissions as their relata, one tests for stability by assessing how many worlds Lewis's first counterfactual is true in. Hopefully it's obvious that (3a) would fail to be true in most nearby worlds. For even in those close worlds where I am not abducted by aliens, there are many relatively small changes that could stop me drinking the coffee — I might not like coffee or desire it at that moment, I might desire the coffee but lack the motivation or resources to make it, and so on. In a lot of not too far-fetched scenarios in which I am not abducted by aliens, I don't drink the coffee. These observations suggest that while (3) expresses a true causal relationship, the relationship is nevertheless a relatively unstable one.

5.10.2 Putting the View to Practice

Having outlined the notion of causal stability, let me demonstrate the view by returning to cases discussed in previous Sections. As a reminder, here's the account:

> **Omissive Causal Responsibility:** an agent is causally responsible for some outcome when they fail to act, if (i) the outcome counterfactually depends on the

> omission and (ii) the relationship between the omission and the outcome is causally stable.

First let's consider what the view says about **Driving:**

> Susan is driving in a suburban neighbourhood. Her hands are on the steering wheel when she observes a pedestrian (Megan) crossing the street approximately 100 feet in front of her. She knows she could easily move the wheel to avoid hitting Megan without veering outside of her lane. Instead, she deliberately remains motionless. Her car hits and kills Megan.

As before, we want an account that recognises Susan, and no one else, as causally responsible for Megan's death. Does my view vindicate this judgment? Yes. To see first that Susan is causally responsible, we need to check for causation via the counterfactual test. As we know, Susan's failure to turn the steering wheel makes her a cause of Megan's death because there is counterfactual dependence between the death and Susan's failure to turn the steering wheel. Next, we assess whether the causal connection is stable. This entails asking how many worlds the associated counterfactuals would hold in:

> *If Susan were to fail to turn the steering wheel, then the car would hit Megan*

Evidently, this claim would be true in many close-by worlds. Once we keep fixed the fact that Susan does not manipulate her steering wheel, Megan's death will occur in a range of none far-fetched scenarios. There are only a handful of none far-fetched scenarios I can imagine where Megan continues to survive, even though Susan fails to turn the wheel. For example, worlds where, for whatever reason, the car is moving at a slower speed than in the actual world, so that a collision would result in non-lethal injuries. Or worlds where Megan sees the oncoming car in enough time to move out of its way. Still, barring these kinds of cases it's difficult to imagine non-far-fetched scenarios in which Susan does not turn the car but Megan remains unharmed. We thus get the impression that the causal connection between Susan's omission and Megan's death is relatively stable. Hence, Susan satisfies both condition (i) and (ii) of the account, making her causally responsible.

Next, let us see how the view also rules out seemingly irrelevant omissions, such as Ali the backseat passenger, as forming the basis for causal responsibility. Ali meets condition (i) because, as we saw in Section 7, the death counterfactually depends upon his omission; had Ali turned the steering wheel, Megan would not have died. However, Ali does not meet

condition (ii) because the causal relationship between Ali's omission and the outcome is a relatively unstable one. We can see this by examining how many worlds the associated counterfactual would hold in:

> *If Ali were to fail to turn the steering wheel, then the car would hit Megan*

There are many scenarios where Ali fails to turn the wheel, but Megan nonetheless survives. Some of the cases will include ones previously mentioned; for example, even given Ali's omission, Megan might survive had the car been travelling at a slower speed. Importantly, however, in addition to these scenarios, the causal connection between Ali's omission and the outcome breaks down in circumstances where Susan actually fulfils her moral obligation to manipulate the wheel. Given that this would presumably happen in most close-by worlds, we get the impression that although there may well be a causal relationship between Ali's omission and Megan's death, the relationship is too unstable to constitute omissive causal responsibility.

And the same reasoning can be applied to other seemingly irrelevant omissions mentioned previously. For instance, Wonder Woman's failure to turn the steering wheel does not make her omissive causally responsible on this view. For whilst it's true that Wonder Woman's failure is a cause of Megan's death, the causal connection is an extremely unstable one because in most close-by worlds where Wonder Woman does not manipulate the steering wheel, Megan would survive anyway since Susan turned the wheel.

In addition to getting it right in cases like **Driving** where we want to assign moral responsibility for bad outcomes, the view gets it right when we're looking to assign moral responsibility for good outcomes. In this regard, it does better than the normative analysis. To see this, consider **President**:

> In the midst of a war, the President is strongly encouraged by her special advisors to authorise a nuclear attack on country X. The President rejects their advice, and decides not to authorise the nuclear attack, as a consequence, the 50 million residents of country X remain unharmed.

I said that the normative analysis is unable to account for the fact that the President is causally responsible for the 50 million residents remaining unharmed, because her omission is not the type of event, which, had it occurred, would normally prevent the effect from occurring. The present account, however, is able to accommodate for the intuition that the President's omission makes her causally responsible for the outcome. Firstly, the President satisfies condition (a)

since there's counterfactual dependence between the President's failure to authorise the attack and the safety of the residents. And the omission satisfies condition (ii) because the causal connection between the omission and the effect is a stable one. We can see this by considering the associated counterfactual:

> *If the President fails to authorise a nuclear attack on country X, then the 50 million residents of country X remain unharmed*

This is an incredibly stable claim. Barring a climate catastrophe or some other significant global event, it's difficult to imagine a scenario in which the President does not authorise a nuclear attack but 50 million residents are harmed all the same. So, in virtue of the fact the President's omission meets both condition (i) and (ii), the view says that the President is causally responsible for the good outcome.

To illustrate the plausibility of this view once more, consider a final example we have not yet encountered which is discussed by Woodward (2006). Suppose that you do not donate £50 to a famine relief organisation. X who lives in country A dies of starvation, and X would have lived if you had sent the money. According to the counterfactual test, your omission to donate £50 makes you causally responsible for X's death, since had you sent the money, X would not have died. Many will be reluctant to accept this claim, and again, I think we can trace these reactions to the fact that the causal relationship between X's death and your omission is relatively unstable. Even supposing that you omit to send the money, any number of things could have occurred that would ensure X's survival. As Woodward notes, if the corrupt dictator who runs country A had stolen less foreign aid, if the food transportation network in A had not been disrupted by war, if X had not been weakened by previous malnutrition, and so on, X would not have died (p. 28). In this sense, the causal link between your failure and X's death is comparatively unstable, thus, you are not causally responsible for X's death.

5.10.3 Proximate Causes

One thing to note is that stability is not just another name for how proximate a causal relationship is. For one thing, it is perfectly possible for a distal relationship to be relatively stable given the right relationship between the intermediate causal steps. For instance, the causal relationship between my birth and my death is maximally stable — there's no alteration to the background circumstances which would mean my birth would not, at some point, be followed by my death. As well as being stable, it is also relatively distal (hopefully), or at the very least it's true that my birth is not a proximate cause of my death.

Although, degrees of stability do not *necessarily* equate to degrees of proximity, it stands to reason that, by and large, the more distal the cause event, the more likely it will be that the connection between the cause and the effect will be unstable. This is because there will be more intermediate steps and hence more chances for the causal connection to breakdown following changes in the background conditions. Suppose that we have chain of causal relationships C_1, C_2, C_3, C_4,… C_n which hold in background circumstances B. Suppose that C_1 and C_2 would fail to hold in some set of circumstances B_1, and that C_2 and C_3 would fail to hold in circumstances B_2, and that C_3 and C_4 would fail to hold in circumstances B_3, and so on. Then the causal chain will be disrupted if any one of these other circumstances, B_1 or B_2 or B_3, holds. Hence, the overall stability of the chain C_1-C_n will be less stable that any individual link within the chain. As a general rule then, the more distal the cause the less stable the causal connection will be.

The fact that degrees of stability loosely tracks degrees of proximity is one benefit of appealing to facts about stability to attribute omissive causal responsibility. As Michael Moore (2009) notes, we tend to believe that the more remote the cause the less one can be held causally responsible; one's causal responsibility intuitively peters out over time. Caesar's crossing the Rubicon may well be a necessary condition for my writing this Chapter, in the sense that that writing the Chapter counterfactually depends upon his crossing, but so many other events have also contributed to my writing this Chapter that Caesar's causal responsibility has long since fizzled out (p. 102).

5.10.4 Degrees of Causal Contribution

Another potential virtue of appealing to the notion of stability regards its capacity to capture *the extent to which* an agent is omissive causally responsible. One might think that omissive causal responsibility is the kind of thing that admits of degrees in the sense that when multiple agents bring about an outcome those agents can be more or less omissive causal responsibility relative to their fellow omitters.

It's certainly very natural to think that causation more generally comes in degrees. For one thing we often talk and think as if causation is a graded notion: we wonder about degrees of "causal potency", "causal contribution", "causal efficacy"; something being "the", "main", "chief", or "principal" cause of an outcome; something being the "stronger/weaker", "more/less important" cause of an outcome. For instance, one might wonder whether the principal cause of Tom's rude outburst was his hunger rather than his character, or we might

think that the tomatoes causally contributed more to making the soup than did the pinch of salt. Alongside our ordinary discourse, a graded notion of causation underpins many serious scientific, legal, political and philosophical thesis. For example, a biologist might claim that genetics cause the prevention of heart disease more so than a healthy diet. A historian might argue that Wilhelm II's invasion of Belgium was more of a cause of World War I than was the assassination of Austria's Archduke, Franz Ferdinand. And consider the very live debate among experts from various scientific fields on what causal factors contribute more to the spread of Covid-19. All these statements convey the thought that some events make a larger causal contribution to the occurrence of an outcome than some other events.

In addition to causation more generally it's reasonable to suppose that omissive causal responsibility comes in degrees. To support this idea, consider the following case:

Committee: C_1, C_2, C_3, C_4 are members of an executive committee of a manufacturing company. Every committee member has one vote each, except C_4, the chair of the committee, who has three votes. The committee is asked to vote on whether to replace the company's outdated equipment, a majority of at least three votes is required to enact the policy. On the day of the vote none of the committee members show up, the equipment later malfunctions, injuring an employee.[52]

This scenario has an interesting causal structure because the injury is caused by an indeterminate plurality of events. Still, intuitively C_4's failure to vote to replace the equipment makes them more omissive causally responsible for the employee's injury than do the failures of C_1, C_2, and C_3. C_4's weighted vote means they had more causal power to wield, plausibly making them more causally responsible when they omit to use that power. Stability can make sense of this judgement. Let's suppose that had the members shown up to vote, they would have voted to replace the equipment. There are a total of seven votes, and the minimum votes needed for a majority to pass is three. This means that there are five different combinations of events that will make a sufficient set for the committee to update the equipment. The sets are:

$[C_1, C_2, C_3]$, $[C_1, C_4]$, $[C_2, C_4]$, $[C_3, C_4]$, $[C_4]$.

C_4 makes an appearance in four of these sufficient sets, whereas C_1, C_2, and C_3 make an appearance only twice. If we assume that these kinds of scenarios are the relevant kinds when

[52] This example and the subsequent discussion is based on a case given by Alex Kaiserman (2017, p.4)

measuring stability, then there a more alternative scenarios in which C_4's omission is a cause of the employee's injury, making their omission more stable than the rest.

Another way of putting this is to say that there are more relevant worlds in which the following counterfactual is true: 'If C_4 had failed to vote to replace the equipment, then it would be the case that the employee would get injured', compared to this counterfactual: 'If $C_{(1\text{-}3)}$ had failed to vote to replace the equipment, then it would be the case that the employee would get injured'. Thus, the causal connection between the employee's injury and C_4's failure to vote is the most stable relative to other committee members. In this way, stability appears to track the extent to which an agent can be held omissive causally responsible for some outcome.

Several authors working on causation have claimed that causation comes in degrees,[53] I won't make a sustained argument in favour of thinking that causation or omissive causal responsibility does come in degrees. I only wish to point out that if it does come in degrees, the notion of stability seems to be a suitable candidate for capturing this idea. And this doesn't just go for omissive causal responsibility but causal responsibility more generally.

5.10.5 Relevant Worlds for Measuring Stability

In this Section, I'm going to get a little more precise about the sorts of changes we ought to be invoking to determine stability. Getting specific about which worlds are relevant for assessments of stability is important for two chief reasons. Firstly, the types of changes we apply will obviously have a decisive role in determining whether a causal relationship is stable or unstable. If we only consider worlds in which there's no gravity, for instance, then the majority of causal relationships that hold in the actual world would not hold in these other worlds, and as a result, all causal relationship will be unstable relationships. In terms of causal responsibility, this would mean that no one could be omissive causally responsible for anything. Secondly, the range of worlds we appeal to will influence stability assessments. If we were to quantify over an infinite number of worlds with an infinite number of changes, then every actual causal relationship will be an unstable causal relationship, since the causal connection will not hold in the majority of these worlds. Again, in terms of causal responsibility, this would mean that no one could be omissive causally responsible. In light of

[53] For example, Matthew Braham and Martin van Hees (2009), Robert Northcott (2013) and Huzeyfe Demirtas (2022).

this, delineating the types and range of worlds relevant to measuring stability is crucial if we wish to avoid unsound attributions about omissive causal responsibility.

Thus, far I've been operating on the assumption that the changes relevant for assessing stability are those instantiated in worlds that are close-by. Woodward places a similar constraint on his analysis of the concept:

> At a very general level, the assessment has to do with whether the relationship of interest (understood either as a causal or a counterfactual claim) would continue to hold under changes that do not depart too much from the actual state of affairs or that do not seem too far-fetched (2006, p. 11).

Limiting neighbourhoods to those that are close-by is a conventional restriction placed on most philosophical accounts that employ a modal framework, and for good reason — restricting investigations to close-by worlds goes some way in avoiding erroneous conclusions about the phenomenon we want to capture. In terms of assessing stability, we would not be entitled to quantify over an infinite number of worlds since the majority of these worlds would be faraway. Furthermore, seemingly irrelevant worlds will, by and large, fall outside of our investigation; worlds with no gravity, for instance, are not relevant for measuring stability given that no-gravity-worlds are comparatively faraway worlds.

With that said, this constraint, taken on its own, doesn't seem to provide sufficient guidance for selecting the relevant kinds of changes for assessing stability. The problem is that, in practice, the constraint only sets parameters around what sorts of changes ought to be excluded; it entails excluding departures which are too dissimilar or improbable from an actual perspective. But it doesn't offer any guidance as to which departures ought to be included in our evaluations. For surely not every close-by world will be relevant for measuring stability. What more can we say about the sorts of changes that matter for establishing stability? To help answer this question, we can turn to Woodward again. He notes that "the specific sorts of changes that are regarded as particularly important for the assessment of [stability] may depend on subject matter or disciplinary specific considerations" (2006: 13). By this he means that the reasons for which we have launched our inquiry will influence the changes we take to be relevant for our assessments. Out of those close-by neighbours, the relevant set will depend upon why we're investigation stability in the first place. To illustrate, suppose that a doctor fails to administer medicine to a seriously ill patient in their care, and shortly after the patient dies. Suppose also that we want to understand who is morally responsible for the death. Given

that our subject matter is ascribing moral responsibility, it would be appropriate to consider changes in the circumstances that reveal pertinent information for assigning or excusing moral blame. For instance, we might consider whether, given the doctor's omission, the patient would have died had the hospital ward not been so severely under-staffed, or had the nurse doing the rounds not been suddenly distracted.

Now imagine that we are not looking to assign moral responsibility, but instead we're pathologists investigating what type of illness caused the death. As pathologists, we won't be interested in changes that reveal information about who's to blame, instead, the changes that interest us will be those that indicate the presence and severity of disease in the patient. Hence, it might be helpful to consider, given that the doctor omitted to act, whether the death would still have occurred had the illness been present but in a milder form, or had the patient not been weakened by an existing illness. The general idea is that the purpose for which we want to test for stability should influence the changes we regard as relevant. This is useful guidance; it doesn't merely tell us what sorts of neighbourhoods ought to be excluded, it gives us a departure point from which to start mapping out the sorts of worlds that should be included. [54]

5.10.6 Answerability and Causal Stability

With the account explicated in detail, it should hopefully be clear how an appeal to causal stability can vindicate our judgements about omissive causal responsibility. Before concluding, I want to offer up one explanation for why this might be the case. The answer has to do with the apparent connection between stability and agency. I've argued that of the countless number of nondonigs that happen at every moment only those nondoings for which you are answerable make us causally responsibility. It seems to me that these answerable omissions also happen to be omissions which enjoy a stable causal connection to their effects. This is to say, that by and large, when a person causes some outcome by omitting in a way that expresses her agency, the causal relationship between the omission and the outcome will be a stable one.

[54] I recognise that while one might accept that the relevant changes depend upon subject matter specific considerations, one might worry that such considerations are multifaceted and difficult to make precise, making pinning down the relevant changes a tough task. For those concerned by this, I want to note that getting precise about how to quantify over possible worlds remains a standard problem for those philosophers who endorse Lewisian modal theories to explain a certain phenomenon, and it's not a problem that I can solve here. But let me say that I think such precision might be illusory in any case; the very nature or possible world semantics renders supplying a more exact quantification a near impossible task, and those philosophers who defend a possible world framework might have to accept this implication.

It's no real mystery as to why stable causal relations are relations that exhibit a person's agency. Stable causal connections do not require a particular set of background conditions to obtain in order for the effect to occur, they hold across a diverse range of circumstances. This suggests that the effect's occurrence is not a result of the particular set of circumstances in which the agent finds herself, but rather the effect is primarily a product of what the *agent* does or does not do. Another way of putting this is to say that stable causal relationships are relationships that are *authored by the agent*, as opposed to being authored by the particular circumstances in which the agent finds herself. This suggests that effects produced from stable causal relationships are manifestations of agency.

To be clear, I do not take answerability to be extensionally equivalent to stability, rather I see stability as a way of capturing the concept answerability. In this way, stability is being used as an epistemic framework to ascertain the target of theories of omissive causal responsibility; namely, answerable omissions.

5.11 Conclusion

In this Chapter my chief aim has been to supply an account of causal responsibility that allows us to hold agents morally responsible for their failures to act. I began by arguing that one should not derive attributions of omissive causal responsibility solely from the counterfactual test for causation, since doing so allocated omissive causal responsibility to irrelevant nondoings, that is, those that are not reflective of agency. I then considered whether the normative analysis of causation could fare better as an account of omissive causal responsibly. Although more promising than the simple counterfactual test, the normative analysis also failed to establish a causal link between the omissions for which we are answerable.

Although I only looked at two theories of causation, their inability to vindicate our judgements of causal responsibly suggest to me that they are not trying to capture the same phenomenon that I aim to capture in this Chapter. To put it another way, the discussion suggests to me that causation is a distinct concept from the concept of omissive causal responsibility (and causal responsibility more generally). Unlike metaphysical causation, the concept of causal responsibility is essentially bound up with features to do with agency and answerability.

I then set about defending an account of omissive causal responsibility that is able to capture these features. My view supplements the counterfactual test with an appeal to the notion of causal stability. The guiding principle is that an appeal to facts about causal stability would restrict allocations of omissive causal responsibility only to those nondoings that are reflective

of a person's agency, and therefore, have to potential to ground moral responsibility assessments.

CHAPTER 6

Moral Praise, Right Reasons and Causal Robustness

In the previous Chapter, I argued that facts about causal stability partially determine whether one is morally responsible for one's failure to act. In this Chapter, I'll explore another way in which causal facts determine moral facts. This time the types of moral facts I'm interested in concern assessments about an action's moral praiseworthiness, or as it's referred to in the Kantian literature 'moral worth', and the types of causal facts I'm interested in concern facts about the causal robustness of an agent's motive. I'll argue that whether and to what extent one is praiseworthy for doing the right thing depends upon the causal robustness of one's motivation.

6.1 Introduction

According to a popular approach to moral worth, a right action is worthy of praise if and only if the agent performed it in response to the relevant moral reasons, that is, the reasons making it right. Call this the Right Reasons Thesis (RRT). The central idea behind this doctrine is that moral worth is not about doing something right because it is right, rather it is about doing something right for the reasons which make it right. This Chapter has two primary ambitions. The first is to show that RRT is not as successful as contemporary discussions suggest. This is because the view fails to adequately satisfy two important desiderata associated with theories of moral worth:

1) DEGREES: A theory of moral worth ought to successfully identify the extent to which an action is praiseworthy.
2) OVERDETERMINATION: A theory of moral worth ought to identify if right actions produced from overdetermined motives have moral worth.

The second ambition of this Chapter is to demonstrate that RRT can satisfy the desiderata when the theory attends to certain causal facts; specifically, the causal robustness of the agent's motive. Broadly speaking, causal robustness concerns the extent to which an agent's motive would continue to produce right action in counterfactual circumstances. I'll say more about the notion of causal robustness shortly, for now the basic idea is that by aggregating the number of counterfactuals the agent is motived to respond to the right reasons in we can determine the

causal robustness of her motive; the more counterfactuals in which the agent continues to have a praiseworthy motive in the more causally robust that motive. I argue that it is in virtue of attending to causal robustness that the proposal is able to satisfy the above desiderata. Let us call RRT combined with a causal robustness criterion the Causal Right Reason Thesis (CRRT).

To clarify, my aim in this Chapter is not to defend RRT, rather my aim is to argue that if you're already an advocate of RRT, then you have strong reasons to adopt CRRT. Not only does an appeal to causal facts provide a more successful theory in virtue of better satisfying the desiderata, it does so in a way that is uniquely unified, intuitive and otherwise theoretically unproblematic.

Roadmap: In Section 2 I clarify the core concepts being deployed in this Chapter — moral worth and causal robustness. In Section 3, I introduce RRT and CRRT in more detail. In Section 4, I outline well-known extensions to RRT which aim to capture degrees of moral worth; I argue that these extensions generate implausible conclusions. Following this, I show that CRRT generates more intuitive conclusions about degrees of moral worth, and hence, better satisfies the first desideratum. In Section 5, I argue that RRT problematically implies that all motivationally overdetermined actions are worthy of praise. I then demonstrate that CRRT is committed to a different nonproblematic claim which better satisfies the second desideratum. Finally, in Sections 6, 7 and 8 I respond to putative objections to CRRT by delineating the account in further detail.

6.2 Clarifications

Before getting started, I'll outline in more detail the two central philosophical concepts being deployed in this Chapter — moral worth and causal robustness.

6.2.1 Moral Worth

Moral worth can be defined as a particular way in which we find an action valuable. It is thought that this value is largely derived from the agent's motive for acting. The moral worth of an action then should not be identified solely in terms of whether the action is a right or good action. Nonetheless, an action's moral worth will likely depend upon whether it is a right action or not. Since I don't want to commit myself to any first order ethical theory here, I'll use the term 'morally right' for acts we would intuitively consider as such, and I use it broadly as to include actions which are obligatory as well as supererogatory.

Not only does moral worth simpliciter rest on whether the action is a morally good action, degrees of moral worth might also depend upon the degree to which the action was a good action. A supererogatory action might deserve more praise than a morally required action merely in virtue of its supererogatory nature. Perhaps the more good the action manifests the more praise it deserves regardless of the agent's motivational profile. This line of thought has not been explored in the philosophical literature (as least as far as I'm aware), still it's worth bearing this thought in mind later when I propose an account for accommodating degrees of moral worth.

Finally, it's important to clarify that I am offering an account of moral worth, or as it's also referred praise*worthiness*. I am interested in when actions are genuinely commendable given that they reflect the agent's will in some substantive way. Considerations of praiseworthiness can be independent of whether we have grounds to treat or react to a person a certain way given that they are praiseworthy.

6.2.3 Causal Robustness

Causal robustness is a relatively underexplored concept in the philosophical literature. Recently though the term has been used in the philosophy of biology (Raerinne 2013, Irvine 2015, Huneman 2010), and it's appeared in some empirical research exploring the ways in which our causal concepts influence our pretheoretical moral evaluations (Grinfeld et al. 2020). Within this research the concept of causal robustness is not especially well-defined and different authors understand the concept in slightly different ways. In this Chapter, I understand robustness roughly as: a causal relationship is robust to the extent that the cause event occurs thereby producing the effect across a range of counterfactual scenarios. The more causally robust the relationship, the more counterfactual scenarios in which the cause event occurs thereby producing the effect. Conversely, the more fragile the causal relationship, the fewer counterfactual scenarios in which the cause event occurs thereby producing the effect. Hence, a maximally fragile causal relationship is one where the cause event occurs in the actual circumstances and thereby produces the effect, but the cause event would fail to occur in any departure from the actual circumstances thereby failing to produce the effect.

One important thing to note is the distinction between the notion of causal robustness and the notion of causal stability which I discussed in the previous Chapter. These concepts are similar insofar as they are concerned with whether a causal relationship would continue to hold in departures from the actual circumstances. However, the two concepts are importantly different.

When checking for stability our aim was to identify whether the outcome or effect would continue to occur in different circumstances when the cause occurs. The check for stability therefore entailed *holding the occurrence of the cause event fixed across alternative circumstances*. We asked: 'Would the effect occur given that the cause occurred?'. When checking for robustness, however, our aim is to identify whether the cause would continue to occur and thereby enter into a causal relation with the effect in circumstances that depart from the actual situation. As a result, measuring for robustness requires that *we do not hold fixed the occurrence of the cause*. Instead, we wiggle various sets of background conditions to check whether the cause event would continue to occur. So, although robustness and stability are both concerned with what would happen to a causal relationship under departures from actuality, their respective tests measure different features of the causal relationship. They therefore offer up different kinds of causal information.

Another way to draw out the distinction between these two concepts is by seeing that a causal relationship can have one of these features but not the other. For instance, the causal connection between certain mutations in the F8 gene and haemophilia is relatively stable. These mutations in the F8 gene will cause someone to contract haemophilia even if we intervene by varying a wide variety of background conditions. But the causal connection between mutations in the F8 gene and haemophilia is not a particularly robust one. Very specific background conditions are needed for these specific F8 mutations to occur in the first place — the cause event occurs only in a handful of background circumstances. Hence the relationship between mutations in the F8 gene and haemophilia is causally stable but not causally robust. Once certain mutations to the F8 gene have occurred, there are very many circumstances in which a person would contract haemophilia, but there are only very few circumstances in which certain mutations in the F8 gene will occur in the first place.

In this Chapter, I'm interested in establishing the robustness of a specific kind of causal relationship; namely, that between agential motivation and right action. Following the definition outlined above, if one's motive manifests itself thereby causing right action across a range of counterfactual scenarios, then it is relatively robust, whereas if it fails to manifest itself thereby failing to issue in right action across counterfactuals then it is relatively fragile. When tracking the causal robustness of a motive then, we're essentially asking: 'in how many different circumstances would the agent's motive manifest itself consequently causing the same right action?' For simplicity, in this Chapter I'll be mostly assuming that the causal connections between agential motives and right actions are causally stable, which is to say that once an

agent manifest a praiseworthy motive to a sufficiently robust degree, that this will issue in right action. With clarifications out of the way, I turn now to outline the Right Reason Thesis of moral worth.

6.3 The Right Reason Thesis and the Causal Right Reason Thesis

Prominent defenders of the Right Reason Thesis (RRT) include Nomy Arpaly and Julia Markovits. Arpaly proposes that:

> [F]or an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons, that is, the reasons making it right. (2002, p. 226)

Similarly, Markovits writes:

> [M]y action is morally worthy if and only if my motivating reasons for acting coincide with the reasons morally justifying the action — that is, if and only if I perform the action I morally ought to perform, for the (normative) reasons why it morally ought to be performed. (2010, p. 205)

Thus, an agent is praiseworthy so long as she is motivated by considerations that explain why her action is right — it relieves suffering, respects personhood, increases welfare, and so on. Given that RRT finds value in being motivated by the reasons that make something right, the view is often presented as a rival to Kantian accounts which, by contrast, find value in being motivated by rightness *per se.* Arpaly and Markovits object to Kantian accounts on the grounds that they are unreasonably restrictive. To illustrate their complaint, consider the now familiar case of Mark Twain's Huckleberry Finn. Huck regards slavery as a legitimate form of ownership, he consequently feels tremendous pangs of guilt when he lies to the slave catchers about the whereabouts of Jim, a runaway slave, thereby securing Jim's freedom. In doing what he believes to be the wrong thing, Huck is not motivated by the rightness of his action, still, it seems like Huck is praiseworthy, and further, his praiseworthiness can be explained by the fact that his helping Jim is driven by a response to the relevant moral reasons — a recognition of Jim's personhood.

RRT has attracted many contemporary sponsors.[55] I suspect that a large part of the account's appeal is its ability to accommodate for the *non-accidentality constraint;* the highly intuitive

[55] For example, Amy Massoud (2016), Errol Lord (2017) and Daniel J Miller (2018).

thought that morally worthy actions are non-accidentally right.[56] Non-accidentality is a central feature of praiseworthy actions recognised by Kant:

> For, in the case of what is to be morally good it is not enough that it *conform* with the moral law but it must also be done *for the sake of the law*; without this, that conformity is only very contingent and precarious, since a ground that is not moral will indeed now and then produce actions in conformity with the law, but it will also often produce actions contrary to law. (1997: 4:390)

Kant rightly notes that morally worthy actions must be issued from a motive that is sufficiently grounded in the right sorts of considerations, otherwise the motive would not be reliable at generating morally right actions. RRT is said to satisfy the constraint because it demands that one ought to perform an action in response to the reasons for which it ought to be performed, thus ensuring a tight connection between motives and morality. The importance of satisfying the non-accidentality constraint cannot be understated; theories are often evaluated in terms of whether they can successfully accommodate for the idea, for if they bestow moral credit upon a wide range of lucky cases we have decisive grounds to reject the view. For instance, if a view were to ascribe praiseworthiness to a person who saves a life only in the hope that their name will be featured in the local paper, then the view ought to be rejected. It would be a mistake to attribute praise to someone who saves a life only because doing so happens to coincide with their self-interested desires.

A second reason for RRT's popularity is entailed by the fact that moral knowledge is not required for moral worth. It doesn't matter if I know the right reasons or if I believe that I am acting for these reasons, all that matters is that *I do in fact* act for these reasons. As a result, people like Huckleberry Finn, who do something morally right whilst believing themselves to be acting wrongly, deserve moral credit (RRT's verdict on the Huck case is often considered a significant virtue of the account).[57]

Despite its strengths, we shall see that RRT lacks the resources to fulfil important desiderata associated with theories of moral worth. Before turning to these, however, I will introduce CRRT, though the introduction will be brief because it will become clear what a fully-fledged account looks like as we go along. For now, I'll say that CRRT maintains RRT insomuch as

[56] I borrow the term 'non-accidentality constraint' from Jessica Isserow (2019).
[57] For discussions regarding the infamous Huck Finn case see, for example, Bennett (1974), Montmarquet (2012) and Sliwa (2016).

praise requires doing right for the right reasons, but unlike RRT, CRRT demands that an agent have a relatively causally robust motive in order to deserve praise. In practice, this entails that the agent not only be responsive to such considerations in the actual world but that they continue to be responsive in a range of counterfactual scenarios. More formally:

> *Praiseworthiness:* For an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons in the actual world, *and* for it to be the case that she would do the right thing for the relevant moral reasons in a range of relevantly similar counterfactual scenarios.

Precisely how many counterfactuals make up a range and what is meant by relevantly similar will be outlined in Section 6. For now, it's enough to say that CRRT turns moral worth simpliciter into a threshold concept whereby agents must clear a threshold by possessing a modest degree of motivational causal robustness which guarantees they will act well in a handful of similar circumstances.

By demanding even a modest amount of motivational robustness, CRRT offers an immediate advantage over RRT — it's more effective at satisfying non-accidentality. As noted, RRT requires a tight connection between motives and morality, and hence, goes some way to securing the constraint. Still in requiring that one's motive be causally robust enough to produce the action in a range of different scenarios CRRT demands an even tighter connection between motives and morality.[58] This virtue might not provide a decisive reason to accept CRRT, but given the importance of the constraint, it does offer an initial motivation for preferring the view. Having now outlined both accounts, I will turn to discuss the desiderata, I begin with DEGREES.

6.4 DEGREES

Sometimes two seemingly morally right actions possess different amounts of moral worth. For example, we may say that Jane deserves more praise for baking her friend a birthday cake than John does for baking his friend a birthday cake on the grounds that Jane is under emotional

[58] Isserow notes that a counterfactual view of moral worth is more successful than accounts like RRT in capturing the type non-accidentality constraint I sketch out here because "non-accidentality seems to require some measure of counterfactual robustness" (2019: 256). Isserow also categorises this type of counterfactual threshold view as a "strong dispositional view" (2019: 254). I do not adopt this label because CRRT does not aim to measure one's overall disposition to act well, rather it aims to measure the robustness of a particular motive.

strain having recently suffered a bereavement. RRT, as it stands, does not discriminate between John and Jane because the view treats all praiseworthy actions as having equal moral worth, yet a comprehensive theory will go further in identifying the *extent to which* an action has moral worth.

In this Section, I look at two different ways in which RRT has been extended to capture degrees of moral worth — the first defended by Markovits, the second by Arpaly and Timothy Schroeder. I intend to show that both of these proposals fail to fully deliver on their promise of satisfying DEGREES. As I see it, Markovits's proposal falls short on two counts; firstly, it doesn't capture the full spectrum of degrees of moral worth, and secondly, it doesn't seem to establish an action's degree of praise*worthiness*. Arpaly and Schroeder's proposal, on the other hand, commits us to unintuitive conclusions about degrees of moral worth. Having raised these objections, I go on to offer CRRT's alternative solution.

6.4.1 DEGREES and RRT

In order to satisfy DEGREES, Markovits combines RRT with an appraiser-relative approach to moral worth. How much praise an action is owed depends upon how we, as appraisers, would have acted had we been in the agent's shoes. In regards to maximally praiseworthy actions, Markovits states:

> A heroic action is a right action (of some moral significance) that most of us, judging the action, would not have had the moral strength to perform, had we been in the hero's place. (2012, p. 297)

Hence, the extent to which an action deserves moral worth is relative to a community of appraisers — the more unusual it would be for that action to occur in one's moral community the more admiration it deserves. But, of course, there can be disagreements amongst communities. I might judge the fireman to be a hero because I would not have risked my life had I been in his shoes, still his colleagues could reject the compliment; 'he's only doing his job' they might protest. Markovits suggests that when this type of disagreement occurs that both of these assessments could be right, and therefore, one can be appropriately described as both a hero and not a hero at the same time (2012, p. 297).

Markovits presents an interesting extension to RRT. However, I worry that it doesn't completely capture the full spectrum of praiseworthiness. The account offers the resources to identify maximally praiseworthy actions — these are actions that most of us would not have

performed ourselves. But what of behaviour that falls short of this extreme? A friend of the account could delineate the proposal further to capture these more ordinary actions. For instance, one could say that if a heroic action is one that most people judging the action would not have the moral strength to perform themselves, perhaps a considerably praiseworthy action which falls short of heroism is one that *nearly most* of the people judging would not have the moral strength to perform themselves. By way of example, if 90% of people consider themselves unwilling to replicate the behaviour of the agent, then the act warrants maximal moral admiration, whereas, if only 70% judge themselves unwilling then the action is certainly praiseworthy but not heroic. This method aggregates moral appraisals to find a kind of mean which then corresponds to the degree of moral worth.

Although initially plausible, it's not obvious to me that an appraiser-relative view is compatible with aggregating judgements in this way. For recall, that if communities disagree about whether an action is heroic or not heroic, then the action itself might appropriately be described as heroic and not heroic. This implies that moral appraisals cannot be combined to generate new moral appraisals; we cannot take a heroic judgement, add it to a non-heroic judgement to get 'almost heroic'. In other words, moral appraisals cannot be aggregated with a view to working out the mean. If this is the case, then the extension seems to fall short of fully satisfying the desideratum, for it only explains when an action deserves maximal credit, leaving a large swathe of more commonplace behaviour unaccounted for.

My second, and perhaps deeper worry, concerns whether the appraiser-relative approach captures praise*worthiness*. On Markovits's view an action is more praiseworthy relative to appraisers, this means that degrees of moral worth depends upon how appraisers stand in relation to the action: namely, whether they would have had the moral strength to perform that action. However, moral worth is typically understood to be a feature of an action that goes above and beyond standing relations. To illustrate, suppose that Wonder Woman performs a dangerous rescue to save a group of children, we don't think her action is made less praiseworthy by the fact that Superman would have performed the same dangerous rescue had he been in Wonder Woman's place. We might think it would be *inappropriate* for Superman to *praise* Wonder Woman by, for example, applauding her after witnessing the rescue, because this gesture comes across as condescending or disingenuous given how Superman stands in relation to the action. But questions about whether it's appropriate to praise someone are

importantly different from questions about whether an action is genuinely worthy of praise.[59] To my mind, the extent to which an action deserves praise is independent of how appraisers stand in relation to the action, and if this is the correct way to think about praiseworthiness, then perhaps the agent-relative approach is the wrong way to determine it.

Some might remain unmoved by this worry, however; one could continue to hold onto the thought that degrees of moral worth is a relativised feature of an action and that the standing relation accurately tracks this feature. Even so, I think taking up this line of argument would be difficult if one were an advocate of RRT. This is because RRT supplies conditions for moral worth simpliciter which do not relativise the phenomenon; whether an action is praiseworthy or not praiseworthy is determined independently of subjective judgements and beliefs at particular times and places. Hence, if one wanted to advocate for both RRT and an appraiser-relative approach, some story has to be told as to why degrees of moral worth is a relativised feature of action yet moral worth simpliciter is not, and in the absence of such an explanation, I think we have reasons to be cautious about adopting the view.

So much for the appraiser-relative extension to RRT, let's now consider a different approach developed by Arpaly and Schroeder. Arpaly and Schroeder take the degree of praiseworthiness to depend on how strongly the action manifests an intrinsic desire for the right-making features. Since strength of desire seems like the type of thing that can be scalar, it easily explains how praiseworthiness can come in degrees: one whose good action manifests a stronger desire for the good is more praiseworthy than one who manifests a weaker desire.

This account may strike you as similar to the one I aim to develop here. In the introduction I stated that CRRT will meet the desideratum by appealing to facts about the causal robustness of the agent's motive. One might initially suppose that strength of desire could be something understood in terms of motivational causal robustness. For example, how strongly one desires the right-making features could be cashed out by identifying how many counterfactual scenarios the agent's motive would continue to produce that right action in. However, there's an important feature of Arpaly and Schroeder's view that precludes it from being understood in terms of causal robustness — it does not attend to counterfactual motives but only the

[59] In recent years, there has been a flurry of work on so-called *standing to blame which involves* identifying facts about the blamer that are relevant to whether an instance of blame is appropriate. There appears to be much less said on standing to praise (with the exception of Kasper Lippert-Rasmussen (2021). Nonetheless, it's reasonable to suppose that, like the relationship between blame and blameworthiness, there will be instances where someone is praiseworthy but praising them is inappropriate given the praiser's standing.

strength of desire which is *actually manifested* in action. To illustrate the idea of actual desire manifestation, Arpaly and Schroeder ask us to imagine two agents who kindly give a lost motorist directions; the first agent is a moral saint with bottomless good will, whilst the second has a quite average amount of good will. For Arpaly and Schroeder, "an opportunity to assist a lost motorist is not typically an occasion for a full display of a powerful commitment to morality. Hence, the strength of the desire for the right or good that is actually manifested in the two cases we imagined is the same" (2013, p. 189). Accordingly, the two agents deserve the same amount of moral credit, despite the fact the first agent generally possesses more good will than the second.

On other occasions, the interior life of the agent can present opportunities to display a powerful commitment to morality, thus making a good action more praiseworthy than it would ordinarily be. This happens, for example, when a person experiencing depression continues to do good despite undergoing great sadness. It takes a strong desire to respond to moral reasons in the grips of depression, and so assuming that this desire is manifested in her actions, "the sorrowing agent is more praiseworthy for her action than a person would be for doing the same good works without having to overcome the same psychological barriers" (2013, p. 189).

Whilst focusing on the strength of desire manifested in action nicely tracks our judgements in these types of examples, I think it fails to do so in other cases, namely, in cases where agents possess a strong competing self-interested desire to act otherwise. To illustrate my concern, consider the following example:

> **Donation**: Two agents, Lola and Kirke, receive a £500 work bonus. Shortly after receiving the bonus, their employer reminds them of the company affiliated charity — UNICEF. Both agents decide to donate their bonuses to UNICEF and both do so for the right sorts of reasons, but Lola and Kirke experience very different internal processes before they come to this decision. Kirke feels a variety of self-interested desires to spend the money on himself, 'after all', he reasons, 'I've earned this money through hard work, I ought to treat myself'. Kirke's desires to keep the money are strong; it takes him a few hours of painful deliberation and pacing before he overcomes his internal struggle and is able to donate. Lola, on the

> other hand, feels no internal resistance or temptation to spend the money on self-interested pursuits. After receiving her bonus, she swiftly gives it to UNICEF. [60]

In **Donation**, both agents perform the same morally desirable action in response to the right sorts of considerations, but we can suppose the strength of concern manifested in their respective actions are different. To put it somewhat artificially, suppose that Kirke's desire for the right-making reasons has a strength of 50 and his desire to keep the money has a strength of 49, while Lola manifests a desire of 40 for the right-making reasons and has no self-interested competing desire. Given that Kirke's donation manifests a stronger desire for the good, an account, like Arpaly and Schroeder's, that posits strength of desire manifestation as a criterion for degrees of moral worth would determine that Kirke deserves *more* moral credit than Lola.

I think that this conclusion is too quick. For one thing, it seems obvious that Kirke should not receive special admiration just because he eventually managed to resist temptation, to argue otherwise would be to penalise Lola for lacking such temptations in the first place. Moreover, praising Kirke more than Lola would be especially dubious if we think desires are the types of things we can have agency over. If we have the power to regulate and reform our desires, then a person who finds it difficult to do well because they have failed to appropriately govern their self-interested desires, should not, other things being equal, be given more moral credit compared to a person who finds doing well easy as a result of the fact they've effectively regulated their desire profile.

A defender of the desire manifestation view might be tempted to debunk my intuition that Kirke is not more praiseworthy than Lola by appealing to judgements about Lola's character. The response might go like this: the fact that Lola donates her money with ease provides evidence to suggest that she's a good *person* and this thought distorts evaluations of how much praise she deserves for her *action*, namely, it leads us to think that she deserves more praise than she actually does. This would be problematic because we would be assigning value not only to the good motive but also to the feature which makes it easy for her to act on this motive — her character. Hence, we would be conflating moral worth with moral virtue.[61]

---

[60] This example is one adapted from Kelly Sorensen (2010), and like Sorensen, we should imagine that Lola and Kirke share similar economic circumstances. £500 is not a trivial amount of money for them, but equally, forgoing the bonus will not deprive them of any necessities.

[61] The objection that a counterfactual account of moral worth tracks moral character instead of moral praise has been advanced by many including Herman (1981), Markovits (2010) and Isserow (2019).

Although it's important to bear in mind these distinctions, I don't think this explanation debunks the targeted intuition because the intuition does not rest on a mistaken conflation between types of moral appraisals. We can see this by filling in the details of the case in a way that makes it clear that Lola's score on the character dimension of appraisal does not unduly inflate her score on the moral worth dimension of appraisal. To do this we simply stipulate that Lola, in fact, has a subpar moral character and that when she donates her bonus she acts out of character. In this story, Lola is occasionally generous, kind, honest, etc., but for the most part, she experiences desires that push her towards the morally neutral, and occasionally the morally bad, she certainly does not typically display the kind of generosity required to donate £500 to UNICEF. Despite this, on the day her bonus arrives, she forms an uncharacteristic urge to relieve the suffering of those less fortunate, thereafter she donates the money with ease. Although I've now specified that Lola's character is somewhat substandard, I take it that our judgements about how much moral credit she deserves remain the same; she's just as praiseworthy for giving away her bonus irrespective of whether she's a virtuous person or not. With this line of argument dispelled, we're back to the thought that praising Kirke more than Lola would be an error, thus I think we have reason to believe that the strength of desire manifested in action does not track the degree of an action's moral worth.

6.4.2 DEGREES and CRRT

I think an alternative solution to satisfying DEGREES can be found by taking a step back to consider the fundamental desideratum on moral worth — the non-accidentality constraint. Recall that moral worth simpliciter depends upon whether an action was brought about through luck or accident. If it's accidentally right, then it's not a candidate for moral worth. It's reasonable to suppose then, that the degree of moral worth depends on the degree to which the action depended on luck. One way to capture the degree of luck involved in action is by looking at the causal robustness of the agent's motive. If the agent has a causally robust motive which means she would continue to do well in a broad range of counterfactual scenarios, then certain circumstances were not needed to bridge the gap between motivation and rightness. For such an agent, her praiseworthy motive plays a leading role in generating action, we can therefore be sure that it's her motive and not the environment which is worthy of credit. Whereas, if an agent has a fragile motive and fails to do well in lots of alternative scenarios, then certain circumstances were needed to forge the connection between motivation and rightness in the actual world, thus, her action is dependent more on luck.

To illustrate this thought, suppose that Kirke's good motive is moderately robust; he's able to overcome his self-interested desires and donate to UNICEF in many relevantly similar scenarios. In his case, Kirke's actually doing well is obviously no accident. If, by contrast, his motive was precarious enough such that he would fail to donate in slightly different scenarios, say, in ones where he's hungry, irritable or he forgets his online banking login details, then his action warrants little praiseworthiness, since his actual action seems almost accidental, creditable to his remembering his banking details and his employer's prompt more so than his desire to relive the suffering of others.

I propose then, that DEGREES is solved by attending to how causally robust the praiseworthy motive is, where causal robustness is cashed out in terms of what proportion of relevant counterfactuals the motive would continue to produce right action in. Simply put, the more counterfactual situations one would continue to perform the same desirable action in, the more causally robust the motive, and thus, the more praise one deserves. Conversely, the fewer counterfactual situations one would continue to perform the same desirable action in, the more fragile the motive, and thus, the less praise one deserves. In short, the amount of moral worth awarded is proportional to the causal robustness of one's motive, and how robust an agent's motive is acts as a proxy for something more important — to what extent the action is a product of accidentality.

With this new condition in place, let us take stock of what has been said about CRRT thus far. In Section 2, I stated that CRRT fashioned moral worth into a threshold concept — one must respond to the right sorts of reasons not only in the actual world but also in a range of possible worlds to gain moral worth simpliciter. Combining this with what I've said about DEGREES, it follows that once an agent has met this threshold, we can move to ask how many other worlds she would do well in, the more of these other worlds she would do well in the more praise she deserves.

6.5 OVERDETERMINATION

I use the term overdetermination for cases in which one has two or more independent motives for doing the right thing and would have acted rightly from any one of those motives even in the absence of the others. Had one motive not been present the agent would have acted anyway. The category of overdetermined actions which present difficulties for theories of moral worth are those in which one of the motives is praiseworthy and the other is not praiseworthy. For advocates of RRT, the specific worry will arise when an agent does the right thing for the

reasons which make it right whilst also being moved by reasons which do not make it right. To illustrate, imagine a politician volunteers to help at a food bank, and she has two motives for doing so:

M1: It is in her career interest to be seen volunteering.

M2: She desires to relieve the suffering of those less fortunate.

Supposing that M2 constitutes doing something right in response to the right reasons, are we to say that on this occasion her action was one done in response to the right reasons, and thus, had moral worth?

In this Section, I evaluate some answers to this question. I demonstrate that RRT's answer is problematic because it risks violating the non-accidentality constraint. Hence, by meeting one desideratum (OVERDETERMINATION), RRT violates a different and perhaps more fundamental desideratum. I next show that CRRT maintains a different claim about overdetermined actions, and unlike the claim RRT is committed to, this claim captures OVERDETERMINATION without violating the non-accidentality constraint therefore providing a more successful solution.

6.5.1 OVERDETERMINATION and RRT

As it stands, RRT is committed to something like the following:

> *All*: All motivationally overdetermined actions have moral worth when at least one of the motives was a response to the right sorts of reasons.

What makes a right action morally praiseworthy according to RRT is the fact that the agent responded to the reasons which make the action right, nothing in the account rules out actions as praiseworthy in virtue of the person having additional motives for doing what they do. Consequently, the view entails that all motivationally overdetermined actions are worthy of praise, on the condition that at least one of the motives was a praiseworthy one. Turning to the politician case, RRT would maintain that the politician is praiseworthy for volunteering because at least one of her motives — M2 — is a response to the right sorts of reasons.

Before moving on, it should be noted that as far as I'm aware prominent defenders of RRT have fallen silent on the question of overdetermination except for Markovits who writes in a footnote: "if there are cases of motivational overdetermination, it may be okay to have some nonmoral motivations for doing the right thing, so long as we're also fully motivated by the

actual normative reasons justifying the act" (2010, p. 238, fn. 66). This brief remark gives little guidance on the question at hand other than to indicate that the theory is amendable to the idea that overdetermined actions may be praiseworthy provided that the agent is fully motivated by the relevant sorts of considerations, though it's unclear what being fully motivated entails. In any case, in the absence of any detailed discussion, I think it's fair to categorise the account as endorsing *All.*[62]

In her influential paper, Barbara Herman points out that endorsing *All* is problematic because we would end up praising some actions which are only accidentally right. Here is what she says on the matter:

> As circumstances change, we may expect the actions the two motives require to be different and, at times, incompatible. Then […] an agent might not have a moral motive capable of producing a required action "by itself" if his presently cooperating nonmoral motives were, instead, in conflict with the moral motive. That is, an agent […] could, in different circumstances, act contrary to duty, from the same configuration of moral and nonmoral motives that in felicitous circumstances led him to act morally. (1981, p. 367)

Here Herman argues that when circumstances change, we may expect the two motives that were hitherto compatible to become antagonistic, pulling the agent towards different ends. During such conflict the agent may feel the pull of the non-praiseworthy motive more than the praiseworthy one leading them to act contrary to duty. Attending to the possibility that the agent would *not* act well in the altered circumstances introduces the suspicion that the original configuration of motives produced right action only accidentally. The conditions of cooperation between the two motives which led to right action in the actual situation depended upon the fortuitous alignment of favourable circumstances. These actions are more a function of the accidental circumstances and less of function of the praiseworthy motive, and therefore, to praise such performances is to praise only accidentally right actions.[63]

To clarify the problem, consider our politician again. The politician is motivated to volunteer from a praiseworthy motive, M2, and a motive of self-interest, M1. Is she praiseworthy? Possibly not. The fact that she volunteers in this world from these motives is compatible with

[62] *All* also seems to be in the spirit of what Markovits proposes in her footnote.
[63] Herman deploys her insights about overdetermination to argue that an action has moral worth when the primary motive for the action is the motive of duty. I set aside the wider context of her project to focus on her claim that overdetermined actions can be accidentality right.

the thought that if these two motives were no longer pushing her towards the same end she would fail to volunteer. It's easy to conceive of scenarios that make these motives combative rather than cooperative. Suppose that the politician was instrumental in enacting punishing welfare reforms which caused a dramatic increase in food bank usage. On the day the politician is scheduled to volunteer, she learns that the press no longer intend to publish a flattering story about her good deed, instead they intend to run a story accusing her of being a hypocrite for volunteering at a food bank in light of the fact that her policies made them necessary. Now if the politician would fail to volunteer in a world where she would receive negative publicity for doing so, then it reveals that her doing well in the actual world depended upon her two motives uniting in the way they did, and the reason they unite in the way they did is due to certain contingent circumstances obtaining. When the politician acts rightly in the actual world then, it is not because of a robust praiseworthy motive, but because accidental circumstances which happened to be favourable in producing right action obtained at the time. Thus, when she acts rightly, she does so somewhat accidentally. [64]

In light of this, we have strong reasons to reject *All*. Whilst this claim does allow RRT to satisfy OVERDETERMINATION, it does so at the cost of violating a more fundamental desideratum on moral worth.

Once *All* is dismissed it might be tempting to consider an alternative that says no motivationally overdetermined actions are compatible with praise. Call this claim *None.* To hold this option is to argue that whenever a non-praiseworthy motive cooperates with a praiseworthy motive to bring about action, the mere presence of the non-praiseworthy motive renders that action devoid of moral worth. Kant has often (although perhaps uncharitably) been viewed as an

[64] One might wonder why I am using Herman's test to check for accidentality in overdetermined actions. Herman's test entails that we consider a scenario where the non-praiseworthy motive is combative rather than cooperative. But, as Benjamin Ferguson (2012) points out, another way to test for accidentality is by considering a scenario in which the non-praiseworthy motive is simply absent. Applying this thought to the politician case, one could ask: why consider a scenario in which volunteering would be damaging for the politician's career interest, as opposed to a scenario in which volunteering is neutral with regards to her career interest? (For example, why not imagine a world where the press do not cover the story at all). In response, I note that in overdetermination cases, the agent treats the non-praiseworthy motive as a relevant reason (and not a mere cause) for or against acting in the actual situation. In the actual world the politician takes the fact that volunteering will improve her career as a reason to volunteer. It therefore seems entirely legitimate to consider alternative cases where the non-praiseworthy motive continues to be *present* and not merely cases where the non-praiseworthy motive is absent. Another way of putting it is if the agent in overdetermined cases takes their non-praiseworthy motive as supplying relevant action-guiding reasons, then our test for accidentality ought to include such reasons.

advocate of this view. [65] He seemingly claims that a dutiful act can have moral worth only if it is done from the motive of duty alone i.e., is not overdetermined.[66] The view has thus been heavily criticised for the apparent consequence that it judges a resentfully performed dutiful act as morally preferable to a similar act done with enjoyment. If I help a friend move house because I promised and because I enjoy helping, my good deed warrants no moral credit according to this Kantian view, since my act is not done solely from duty but also from enjoyment. So whilst *None* does fulfil the desideratum, and plausibly does so without violating the non-accidentality constraint, it comes at the expense of our widely held intuitions about moral worth.[67] For this reason, we ought not place an indiscriminatory ban on actions as candidates for praise in virtue of the fact that a praiseworthy and non-praiseworthy motive were each individually sufficient to bring about its performance. Hence, we should also rule out *None* in the search for some better alternative.

The final option available to us I call *Some*: some motivationally overdetermined actions are compatible with moral worth. In particular, the set of actions that are compatible with praise are the ones that do not violate the non-accidentality constraint. In what remains, I will explain how CRRT accurately captures *Some*.

6.5.2 OVERDETERMINATION and CRRT

Let's peddle back. CRRT says that a right action has moral worth if it's performed in response to the right reasons not only in the actual world but also in a range of possible worlds, and to capture degrees we look to see how many additional worlds the agent would act well in. So how does CRRT satisfy *Some*? The overdetermined actions which are compatible with praise are simply the ones that meet the condition for moral worth simpliciter, for in meeting this condition we can be certain that their actually doing well was not the product of accidental

---

[65] More recently, Philip Stratton-Lake (2000) has endorsed *None*. Following Herman, Stratton-Lake argues that we cannot praise all motivationally overdetermined actions since doing so would risk violating the non-accidentality constraint. And further, he finds no plausible way of being able to discriminate between those sets of overdetermined actions which violate this constraint and those which do not. He thus resigns himself to the conclusion that "overdetermined acts cannot, therefore, have moral worth" (p. 108).

66 In the Groundwork for the Metaphysics of Morals, Kant famously says of the man who is so overcome by sorrow that he is no longer moved by the needs of others: "suppose that now, when no longer incited to it by any inclination, he nevertheless tears himself out of his deadly insensibility and does the action without inclination, simply from duty; then the action first has its genuine moral worth" (1997 4:398).

67 Many commentators have sought to amend Kant's proposal in order to avoid this aspect of the theory. For an interesting discussion see Henson (1979), Herman (1981) and Benson (1987).

circumstances which fostered cooperation between the praiseworthy and non-praiseworthy motive — luck could not persist across modal universes in this way. If, on the other hand, the praiseworthy motive was fragile enough to the extent it could easily be overridden by non-praiseworthy motives in most similar scenarios, then the agent's actually doing well was a result of the accidental cooperation of praiseworthy and non-praiseworthy motive, therefore, they deserve no moral credit.

To clarify, consider our politician again. Recall that the problem with praising her for volunteering was the thought that she would not volunteer in a world where M1 and M2 no longer cooperated, that is, in a world where volunteering conflicts with her career interest. What does CRRT say about this case? Generally, it says that the politician is praiseworthy if her moral motive were sufficiently causally robust to see her volunteer in a range of relevantly similar scenarios, but she is not praiseworthy if her motive is sufficiently fragile such that she would fail to volunteer in these scenarios.[68] To find out if the politician's action deserves praise then, we must get precise about what counts as a relevantly similar scenario and how many of these scenarios constitutes a range. I attempt to do this in the next Section.

6.6 Relevant Counterfactuals for Measuring Robustness

According to CRRT, moral worth simpliciter requires possessing a somewhat causally robust praiseworthy motive. I've cashed out causal robustness in terms of counterfactuals such that moral worth simpliciter requires acting rightly in response to the right reasons not only in the actual circumstances but also in counterfactual circumstances. Degrees of moral praise is also determined by whether an agent would have been motivated had things been different. But appealing to facts about causal robustness might strike you as odd insomuch as possessing or failing to possess a praiseworthy motive in some alternative counterfactual doesn't seem to matter to the moral worth of the actual action. Consider the following example:

> Aisha runs a marathon for charity in the actual world, but had she fallen at the start line and broken her ankle, her motive would not have been robust enough to produce right action, and she would have failed to compete in the race as a consequence.

[68] Another way of putting this is to say that the original case is under-described in terms of establishing moral worth. We need further *counterfactual* information about the relative robustness of her motives. I suggest that the under-description retrospectivity explains the competing determinations around the case.

If we maintain that evaluations of moral worth are sensitive to facts about causal robustness, then we might conclude that whatever amount of credit Aisha deserves is mitigated by her failure to do well in the broken-ankle-world. I agree that this would be the wrong conclusion. In response, one might be tempted to reject appeals to causal robustness altogether, but given that non-accidentality seems to require some measure of robustness, I think this move is overhasty. A more amicable solution, and the one I undertake here, is to identify the counterfactuals that matter to measuring causal robustness and those that do not.

For a first pass, we might think that the counterfactuals which matter are those instantiated in nearby worlds — worlds most similar to the actual world. Accordingly, possessing a causally robust motive sufficient for moral worth simpliciter would entail being motivated to act rightly in the actual world and a range of nearby possible worlds, and once this threshold is met the more nearby worlds one is motivated to act well in the more praise one deserves. Privileging nearby worlds has initial plausibility because our moral worth judgements appear to be sensitive to motivational changes that occur in very similar conditions, while they don't appear to be sensitive to changes that occur in radically different conditions. For instance, we seem to care if our marathon runner would be motivated differently had she been unable to wear her favourite running top, because a failure to do well in this world shows that her good motive, while grounded in the right sorts of reasons, was not sufficiently robust. But we don't care how she would be motivated when circumstances are drastically different — we're not interested, for example, what she would do had the marathon taken place in an apocalypse because this world is too modally distant to render any important information regarding her actual action.

Despite its initial plausibility, I think that privileging nearby worlds would be a mistake. This is because nearby worlds occasionally contain circumstances that are significantly more demanding, and when they do the counterfactual test misfires. This happens when a nearby world's slight deviation from the actual circumstances leads to a confounding turn of events which, as a consequence, demands a greater personal sacrifice from the agent to do the right thing than originally expected. In such cases, one could hardly deny that someone's actual action lacks moral worth just because, had the moral stakes been significantly higher, they would be unwilling to perform that action.[69] I think this thought explains why we take a lot of

[69] Markovits has objected to a counterfactual account on these grounds. She states that "we should not think [an action is] *less* worthy because the agent who performs it (still for the same right-making reasons) might *not* have done so had the cost been higher (2010: 213). However, this objection does not show that a counterfactual account of moral worth is incorrect but rather that not *all* counterfactuals are relevant to moral worth.

counterfactuals to be irrelevant to decisions about moral worth. It explains our readiness to ignore Aisha's counterfactual motive in the broken-ankle-world; the broken-ankle-world is a relatively nearby one (not much has to change for us to get there, perhaps the strategic placement of a shoelace), but what unfolds as a consequence makes doing the right thing vastly more difficult, and thus, we take the scenario to have no bearing on the moral worth of what she actually does. The counterfactual test aims to identify if the motive is robust enough to transcend very particular circumstances, it's not intended to identify if the motive is unbreakable.

If the relevant counterfactuals are not those instantiated in nearby worlds, then which ones are relevant? Here is one suggestion. When we decide on an action's moral worth, we implicitly associate the action with a set of conditions that we take to be *normal* for its performance. The same is true of those doing the performing; action guiding-decisions are made in light of our expectations about how things would normally turn out. Aisha's decision to run a marathon for charity, for instance, is made with the reasonable expectation that she will not be required to perform her good deed having sustained a severe injury. In light of this, when we consider how the agent would have acted had things been different, we ought to fix the normal conditions which contextualise the performance of the action. Manifestly, this means considering counterfactual scenarios that are instantiated in worlds that are comparatively normal from an actual world perspective.

Ranking worlds according to their comparative normalcy is not new, although it is not conventional either. We saw in Chapter 1 and 2 that the notion of normality is increasingly being invoked in the metaphysical domain to determine causal facts; causes being those events that deviate from what we would normally expect to occur. Still, determining causal facts by explicitly appealing to a framework of possible worlds ranked according to their comparative normalcy remains unconventional.[70] Outside of the causation literature, Martin Smith (2007, 2010) has explored ordering worlds based upon their comparative normalcy in connection with epistemic justification and ceteris paribus conditionals.

I'll characterise a normal world by employing the notion of normality that I described in Chapter 2. To recap, I said that the notion of normality was constituted by a plethora of more specific norms. These norms are both prescriptive and statistical. To say something is a

[70] Notable exceptions include Peter Menzies (2004) and Joseph Halpern (2016) who both rank worlds according to their comparative normalcy in order to determine causal claims.

statistical norm is to say that it conforms to a statistical mode; for example, if Scotland were to have a rainy winter, the country's weather would conform to a statistical norm. By contrast, to say something is a norm in the prescriptive sense is to say that it conforms to the way things ought to be or are supposed to be; for example, keeping a promise would confirm to a prescriptive moral norm — you're supposed to keep your promises. Broadly speaking then, a world can be categorised as normal to the extent that it abides by actual world statistical and prescriptive norms. A world where Scotland is rainy in the winter is more normal, other things being equal, than a world where Scotland has a dry winter. Likewise, a world where people keep their promises is more normal, other things being equal, than a world where people do not. Abnormal worlds, by contrast, will be those worlds that deviate from actual world statistical and prescriptive norms.

Restricting the counterfactuals that matter to those in normal worlds, has an immediate advantage over privileging nearby worlds — normal worlds typically do not contain significantly more morally demanding scenarios. In normal worlds, circumstances evolve in ordinary ways, in accordance with our expectations, this inhibits confounding changes of events from occurring which in turn prevents drastic changes in the moral stakes. Aisha's broken-ankle-world, for instance, would not enjoy membership in the normal worlds, since breaking one's ankle immediately before running a marathon is statistically abnormal. Privileging normal worlds has another virtue; given that normal worlds share the same norms as the actual world, it appeases the intuition that the scenarios which matter to moral worth are those which are relevantly similar to the actual scenario.

In terms of moral worth then, we're looking to see if the agent would continue to perform the same morally desirable action in alternative circumstances that are comparatively normal from the perspective of the actual world. Having identified the relevant counterfactuals, we are now in a position to specify the necessary and sufficient conditions for moral worth under CRRT:

> *Praiseworthiness:* For an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons in the actual world, and for it to be the case that she would do the right thing for the relevant moral reasons in a range of normal worlds. Once this threshold is met, the more normal worlds the agent would continue to perform the right action in from a response to the relevant moral reasons, the more praiseworthy the action.

With the relevant worlds identified as normal, CRRT can evade a second criticism typically brought against counterfactual accounts. The objection states that in virtue of rendering morally worthy actions counterfactually robust, moral worth becomes too hard to achieve. This point has been pressed by Jessica Isserow and Paulina Sliwa. Isserow writes that "[i]n so far as one's account render's morally worthy actions counterfactually robust, it risks rendering praiseworthy agents far rarer than we take them to be" (2019, p. 258) and Sliwa states: "clearly, it is unreasonable to demand that to have moral worth the agent needs to have acted rightly no matter what. Some contingency must be compatible with moral praiseworthiness" (2016, p. 400). Isserow and Sliwa are right to point out that counterfactual accounts have the potential to be unreasonably demanding, but CRRT escapes this worry. In specifying that only scenarios manifested in normal worlds are relevant to moral worth, the account restricts the number of worlds quantified over thereby limiting the number of scenarios an agent is required to act well in. Precisely how many worlds ought to be included in the range of worlds quantified over for moral worth simpliciter will depend upon how many guarantee non-accidentality. A single counterfactual attempt to act rightly for the right reasons will not guarantee non-accidentality since it too may contain unduly biased circumstances, but sufficiently many attempts scattered across normal worlds will provide a guarantee because not all of these attempts can depend on particularly favourable conditions obtaining. In any case, the threshold will be a moderate one — an agent does not have to do rightly no matter what to gain moral credit. So CRRT does not say you must act well 'no matter what', and it is compatible with contingency. Indeed, CRRT gives contingency its proper place by sampling a range of contingent case, rather than narrowly focusing on just one (which may be an outlier).

6.7 A Potential Worry about Normal Worlds

Before turning to see how my account handles famous examples in the causation literature, I want to face-up to a preliminary objection one might have about using norms to determine moral worth. One might be worried that by deploying notions of normality to determine moral worth, my proposal might run into the same criticism I raised against a normative approach to causal responsibility in the previous Chapter. There I noted that many authors have invoked a normative analysis to identify which omissions, out of our countless nondoings, make us causally responsible for outcomes. Generally, the idea is that omissions which violate normality make us causally responsible for outcomes, whereas omissions which conform with normality don't make us causally responsible for anything at all. I argued that this approach to determining causal responsibility was problematic because, amongst other things, it can lead

to erroneous conclusions about who is causally responsible. For instance, suppose that it's normal for women (and not men) to carry out the domestic labour. When a woman fails to undertake the domestic labour she is causally responsible for any outcome produced by that omission, say, having an untidy house. The problem is that a man who also omits in the same way is not causally responsible for the untidy house because he does not violate a norm. This seemed like the wrong result. If we grant that a woman's failure to do domestic labour makes her causally responsible for the effects of that failure, surely the same causal attributions should be made in regards to her male counterpart. Since I'm also employing notions of normality to determine an agent's moral worth, one might be concerned that CRRT will be committed to similar kinds of erroneous conclusions about whether and to what extent one is praiseworthy. Specifically, one might be worried that, much like the normative approach to causal responsibility, CRRT will get attributions of moral worth wrong when the norms at play are pernicious norms.

However, CRRT does not succumb to this kind of worry; the proposal gets the right kind of verdicts about moral worth even when there are pernicious norms at play. To illustrate consider what CRRT would say about the following case: suppose that Bill and Tamsin live together, and that Tamsin is about to undertake an especially demanding project at work. To ease the pressure on Tamsin, Bill cooks all of their meals and keeps their house clean for a week. But also suppose that in Bill and Tamsin's world it's normal for women (and not men) to carry out the domestic labour. Does CRRT say that Bill is praiseworthy (intuitively it should), despite the fact that Bill and Tamsin's actual world contains harmful norms? Bill's moral worth depends on whether his motive is causally robust enough to produce good action in a range of comparatively normal worlds, and from the perspective of Bill's modal position, normal worlds will be ones where it's normal for woman and not men to do the domestic labour.[71] If Bill is motivated to cook and clean in his actual world where the patriarchal norm already exists, presumably the existence of such a norm in other counterfactual scenarios will not make a difference as to whether he is motivated or not. As such, the fact that Bill does something good in a scenario contextualised by pernicious norms does not detract from Bill's moral worth. To

[71] There is an ambiguity here about whether a world in which women and not men do the domestic labour really is the most comparatively normal world for Bill. In one sense, it would be the most comparatively normal world because it adheres to the actual patriarchal norm. But in another sense, it isn't most normal because it violates a moral norm. It's not quite clear to me how to weigh up such norms to get an idea of what is overall most normal. CRRT would certainly benefit from an analysis of the concept of normality to settle such questions, but sadly is it outwith the scope of this Chapter to provide one.

measure causal robustness we ought to consider whether Bill would be motivated in these patriarchal-norm-worlds when the background conditions are slightly altered; for example, we might ask whether Bill would act well had their house required more cleaning, or had Bill also needed to endure a demanding period at work.

So although the normative analysis and CRRT both deploy notions of normality to determine their target phenomenon, CRRT does not succumb to the same worries that press the normative analysis of omissive causal responsibility. This is no doubt because the two views use normative considerations in their respective framework very differently. The normative analysis is concerned with when an omission violates a single norm, CRRT is not concerned with norm violations but instead with whether a motive would cause right action in scenarios when we keep certain normative considerations fixed.

6.8 Illustrating and Defending CRRT in Further Detail

Having fully specified my account of moral worth, I'll now elaborate the view further by applying CRRT to famous cases in the philosophical literature. But first, let me now return to the politician and deliver a final verdict on the case. According to CRRT, to deserve credit for volunteering, the politician's praiseworthy motive must be robust enough to see her volunteer in a range of normal worlds, we can now ask whether worlds in which volunteering would be against her career interest represent normal states of affairs, and therefore, are relevant to establishing the politician's moral worth. To my mind, it's very plausible that at least some such worlds will be normal. Consider, for instance, the world in which volunteering would gain the politician negative publicity. This state of affairs seems comparatively normal; it's statistically normal for journalists to publish stories ridiculing politicians (far more normal than publishing stories approving of politicians). Furthermore, in running the story the press are abiding by a prescriptive norm in the sense that we believe that the press are supposed to scrutinise the actions of our political representatives. Consequently, we ought to take the fact that she would fail to act well in this scenario as relevant to the politician's moral worth. Even so, considering one relevant counterfactual does not supply enough evidence to deliver a final verdict — we need to consider how she would act across a range of normal worlds. So to settle the case, let us further suppose that the politician's praiseworthy motive is not especially robust; her concern for those less fortunate is not deep or impassioned but fleeting and feeble, as a result, she would fail to act well in many situations where volunteering did not coincide with her self-interest. And if we also suppose that a substantial amount of these scenarios will

contain comparatively normal conditions, then according to CRRT the politician is not praiseworthy.

Let's see how CRRT handles a further two cases much discussed in the moral worth literature. Firstly, consider Markovits's example of the fanatical dog-lover who "performs a dangerous rescue operation to save a group of strangers at great personal risk" (2010, p. 210). Markovits argues that the dog-lover's actual rescue is not made less creditworthy by the fact he would have abandoned the strangers had his dog required his heroics at the same time. Let's assume with Markovits that saving the dog over the strangers would be the wrong thing to do. What does CRRT say about the case? Firstly, we have to determine whether the counterfactual Markovits invokes is a relevant one, in other words, does the scenario in which the dog requires saving at the same time as a group of strangers represent a normal state of affairs? Plausibly, no. Not only would it be statistically unusual for a dog to need rescuing at great personal risk to its owner, it would be even more unusual that this should occur at the very same time a group of strangers also need saving. And I can't see any sense in which the scenario would be prescriptively normal. As a result, there will be very few normal worlds where such a coincidence occurs, and so relatively few cases that have this tension. Thus, it carries no particular weight in determining the moral worth of the dog-lover's actual action.

Finally, consider Isserow's example of the devoted parents. These parents make great personal sacrifices for their children from a concern for their children's wellbeing, but they are so devoted that they would promote their children's wellbeing even when doing so is morally wrong. For example, they may refuse to let their children experience a very small cost in order to substantially benefit many less fortunate children. According to Isserow, the fact that the parents' devotion would produce wrong action in different circumstances does not prevent us from judging their actual sacrifice as praiseworthy; this "strongly suggests to me", claims Isserow, "that judgments of moral praise do not stand or fall with judgments of counterfactual robustness" (2019, p. 263). Does the devoted parent case present a counterexample to the proposal I lay out here? No.

According to CRRT, for an agent to be praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons in the actual world, and for it to be the case that she would do the right thing for the relevant moral reasons in a range of normal counterfactual scenarios. To determine moral worth then, we ask whether that same right action would be performed, for the right reasons, in normal circumstances — the deontic status of the

action is fixed across worlds. Whether and to what extent the devoted parents are praiseworthy depends upon whether they would perform the same sacrifice (conceived of as a right action) for the reasons which make it right in circumstances that are comparatively normal from an actual world perspective. Given that the parents are deeply devoted to their children, evidence suggests that they would continue to make the same right sacrifice in a range of normal worlds, in which case CRRT appeases Isserow's intuition that the parents are praiseworthy.

Notice that the counterfactual test employed by Isserow is different from the one employed by CRRT. Isserow considers whether the agent would continue to perform the action when circumstances make it morally wrong to do so. Whereas, I consider whether the agent would continue to perform the action in circumstances when it continues to be morally right to do so. Since the counterfactual tests are different, the two views produce different conclusions about moral worth. Isserow's complaint, therefore, does not target counterfactual accounts in general, but rather a specific counterfactual test; the objection has no grip on a view like CRRT which does not incorporate that test.

6.9 Conclusion

In this Chapter, I have tried to build on the success of RRT by supplementing it with an appeal to facts about causation. I have used a counterfactual framework as a way to measure the causal robustness of an agent's praiseworthy motive. In this way, counterfactuals have served as an epistemic tool, they have acted as a unifying, reliable and accurate proxy for denoting pertinent information about something which is constitutive of moral worth — motivational causal robustness. The truth value of the counterfactuals in and of themselves I take it is not something that endows an action with moral worth. In any case, I have argued that attending to an agent's counterfactual motives as well as their actual motives, allows us to successfully meet desiderata associated with theories of moral worth. Alongside this, I've argued that invoking counterfactuals means CRRT is able to better secure the non-accidentality constraint — a significant virtue if the problem of moral luck concerns you — and finally, by specifying that the counterfactuals relevant to moral worth are those instantiated in normal worlds, I take CRRT as able to appease the criticisms typically raised against modal accounts.

I wish to outline one final thought. I've argued that moral worth requires that (i) an agent does the right thing for the right reasons in the actual world and that (ii) the agent does the right thing for the right reasons in a range of normal worlds. Thus far I've taken (i) for granted in order to focus on defending (ii), but one might wonder whether taking (i) for granted is justified.

In particular, we might ask whether it is necessary for an agent to do the right thing from a response to the right reasons in the actual world to gain moral credit. Some cases would suggest not. Imagine, that as before Lola is donating her £500 bonus to UNICEF from a praiseworthy motive, however on this occasion, the actual world is not a normal world; moments before Lola clicks the mouse authorising the transaction, part of her ceiling falls in and kills her, she thus fails to satisfy condition (i). But imagine that Lola satisfies condition (ii). She possesses an incredibly robust praiseworthy motive which sees her donate in all normal worlds, including those where she needs to look for her lost debit card, or where she's anxious about a work project, or where she feels tired, grouchy or hungry. In fact, Lola's motive is so robust that she even manages to donate in less normal worlds, say worlds where she's bereaved after the sudden, tragic death of a family member. Despite the fact that Lola performs morally right counterfactual actions from an extraordinarily robust praiseworthy motive, CRRT says that Lola does not deserve moral credit because she fails to act well in one world — the actual world. Some will find this verdict counterintuitive; in particular, if Lola is denied praise, then the conditions for moral worth seem too demanding.

Perhaps this suggests that RRT isn't merely to be supplemented but to be overthrown. We might want to jettison (i) as a necessary condition for moral worth, making it the case that an agent is only required to act well in normal worlds from a response to the right reasons to gain moral credit. Although this will strike many as a radical position, notice that it is motivated in light of a conventional feature of moral worth — the non-accidentality constraint. Given that what happens in the actual world can sometimes depend upon accidents, we ought to doubt whether acting rightly in the actual world is always necessary for an action to have moral worth. More work needs to be done to explore the implications of such a view. But at the very least, cases such as this ought to prompt us to question a central assumption, pervasive in the moral worth literature, that an agent's behaviour in the actual world has a special claim to determining moral worth

CHAPTER 7

Conclusion

The purpose of this thesis has been to show that morality and causation are interdependent. I have argued that (a) facts about causation partly depend upon facts about morality, and that (b) facts about morality partly depend upon facts about causation. For many metaphysicians who think of causation as a mind-independent feature of the natural world, the first of these two claims — (a) — may be alarming. For it entails that causal facts depend upon potentially mind-dependent considerations. The second of these two claims — (b) — is perhaps less controversial. Philosophers largely agree that causation has a determining effect on moral assessments — it is often thought of as the metaphysical glue that connects our conduct to the moral landscape, although there has been much said about what this connection consists of.

Now if we pull back for a moment and consider claims (a) and (b) together, we might find another surprising aspect of the relationship between causation and morality. For if it's true that causal facts depend upon moral facts, and moral facts depend upon causal facts, then we are met with a potential circularity of dependency between causality and morality. In this Chapter, I will conclude with some final remarks exploring and analysing this potential circularity, but before doing so I will first provide a brief review of the critical points made in this thesis.

7.1 Review of the Critical Points

In Chapter 2, I began by introducing two dominant 'meta-causal' views: causal realism and normativism. The former states that causation is a mind-independent structural feature of the world. This view has often been presented as incompatible with normativism, a view which argues that considerations about what is normal, including considerations about what is *morally* normal, play a central role in determining causal facts. The basis for incompatibility turns on the idea that normality is a mind-dependent concept. The causal realist, therefore, is not entitled to deploy the concept to determine causal relations, lest those relations themselves be rendered mind-dependent. I argued that the incompatibility thesis has high-stakes implications for the realist, and for this reason I suggested that the argument for incompatibility warrants more scrutiny than it has thus far received. In this Chapter I therefore set about analysing the arguments given in favour of incompatibility, in addition to exploring how the realist could

resist such arguments. I argued, contra the prevailing narrative, that they are to some extent compatible. I showed that the norms which comprise the notion of normality are not necessarily mind-dependent, some are mind-independent and thus pose no threat to the realist project. My argument opened the door to developing a new restricted kind of normativism which can appeal to a restricted set of normative considerations whilst preserving a realist conception of causation. I then went on to compare this new restricted normativism with an unrestricted kind of normativism — a view which does not aim to secure realist metaphysics, and thus is not restricted in the sorts of normative considerations it can draw upon. I concluded that unrestricted normativism is a more successful approach in virtue of the fact it can better satisfy desiderata associated with theories of causation.

Having clarified the central elements of causal realism and normativism in Chapter 2, Chapter 3 analysed whether James Woodward's (2003) interventionism is compatible with causal realism. My aim in this Chapter was to provide reasons in favour of thinking that interventionism must invoke considerations about normality to determine the appropriateness of a model (these include considerations about what is morally normal). I argued this by taking two widely accepted "realist" criteria for determining the appropriateness of a model — the stability criterion and the serious possibility criterion — before showing that the criteria are essentially bound up in or need to be supplemented with considerations about what's normal and abnormal. Crucially, I showed that the relevant notion of normality and abnormality being deployed here involves an appeal to mind-dependent norms. I therefore argued that a successful theory of interventionism is incompatible with a realist conception of causation.

These two Chapters supported the first aim of the thesis: to show that causal facts depend upon moral facts. Chapters 5 and 6 of the thesis supported the second aim: to show that moral facts depend upon causal facts.

In Chapter 5, I began by arguing that in order to hold an agent morally accountable for the consequences of their failures to act, we need to establish a causal connection between their failure and the consequences of that failure. In other words, we must establish omissive causal responsibility. I outlined two strategies for establishing omissive causal responsibility which were based upon the simple counterfactual analysis of causation and what I called the "normative analysis" of causation. I then argued that the former is too permissive in its attributions whilst the latter is too restrictive, before defending my own account of omissive causal responsibility which incorporates the notion of causal stability. According to my view,

one can only be omissive causally responsible for an outcome when the causal connection between one's omission and the outcome is relatively stable. Throughout this Chapter, I also make the case for thinking that theories of causation will inevitably be found inadequate for establishing omissive causal responsibly (and I suspect causal responsibility more generally) because we want more from our theories of omissive causal responsibility than its ability to merely establish a causal link between omissions and outcomes. We want the theory to establish a causal link between outcomes and those omissions which indicate agency in particular — this being the target phenomenon for theories of omissive causal responsibility. Appealing to facts about causal stability allows my theory to pick out those agency-indicating omissions. This Chapter advanced the debate along several dimensions: it fixed the target and set the success conditions for theories of omissive causal responsibility, and it defended a novel account of omissive causal responsibility.

In Chapter 6 I turned from looking at moral responsibility for outcomes to moral praiseworthiness for actions. I argued that whether and to what extent one is praiseworthy for doing the right thing depends upon causal facts; specifically, facts about the causal robustness of one's motive. I made this argument in the context of responding to a popular theory of moral worth which I called the "Right Reason Thesis". First, I demonstrated that the Right Reason Thesis is not as successful as contemporary discussion suggest because it fails to satisfy two important desiderata associated with theories of moral worth, which I labelled DEGREES and OVERDETERMINATION. Next, I argued that the Right Reason Thesis can meet both desiderata when the theory attends to facts about the causal robustness of an agent's motive. The idea being that facts about causal robustness determine the degree to which an action is praiseworthy as well as whether an action is praiseworthy when produced from overdetermined motives. I thus advanced a new theory of moral praiseworthiness which builds on the success of the Right Reason Thesis which I termed the "Causal Right Reason Thesis".

7.2 A Circularity of Interdependency?

Above I intimated that when we step back from the individual Chapters to consider the overall thesis argument that an ostensible circularity arises. If it is generally the case that causal facts determine moral facts and moral facts determine causal facts, then the dependency relation between causality and morality is a circular one. Some philosophers will find this worrying. Those who subscribe to metaphysical foundationalism, for instance, will regard the non-linear dependency structure as objectionable. According to metaphysical foundationalism, there

exists some absolutely fundamental entities upon which all non-fundamental entities ontologically depend. The various branching strands of ontological dependency must be traced down to some fundamental entities: a collection of things that do not depend on anything, and upon which all the non-fundamental entities above them on the branch ultimately depend. Given that a circular relationship of dependency does not ultimately terminate in some foundational entity, metaphysical foundationalists might find my argument entails an unappealing metaphysical position.

But this kind of worry is unfounded. For the concern to stand, one would have to understand the overall argument as entailing a 'strong circularity'. By this I mean that the circle would be self-contained; moral facts would be *wholly* determined by causal facts, and causal facts would be *wholly* determined by moral facts. This is not what I argue for. In regards to causation, for example, facts about when '*c* caused *e*' depend upon interventions, counterfactual dependence and causal models, and in regards to praiseworthiness, whether a right action has moral worth depends upon the content of the agent's motive. In light of the fact that there are other entities on which these facts depend, there is a possibility that the dependency can take a linear structure which terminates in some fundamental entities.

Even if my argument does not entail a strong circularity, many will think that a weak circularity has unwanted implications in the context of causation and morality. For example, one might worry what this circularity means for the objectivity of moral facts. Take claims about moral responsibility. The received view in philosophy is that causation is an example of a perfectly precise external metaphysical relation (Bernstein 2016, p. 446). Given this rendering of causation, causal facts are thought to be capable of importing a degree of objectivity and precision into moral responsibility claims. But if causal facts and moral facts are interdependent, then we lose a sense in which causation is an external foundation from which we can anchor moral assessments.

But I don't think we should be worried about these sorts of implications just yet, because the argument presented in this thesis does not entail a cyclical relation of dependency between causal facts and moral facts. Thus far I have glossed the potential circularity very broadly between "causal facts" and "moral facts", but in the thesis, I discuss several different kinds of causal and moral facts including facts about: moral responsibility, moral praiseworthiness, moral norms, causal stability, causal robustness, and causal relations of the kind '*c* caused *e*'. Importantly, whether the argument is actually circular depends upon which of these particular

causal and moral facts participate in the determiner/determinant relationship.[72] If the kind of moral facts that determine causal facts are *different* from the kind of moral facts that are determined by causal facts, then there is no circularity for there is no overlap in the kind of facts we're talking about. If, on the other hand, the kind of facts being appealed to are both the determinants and determiners then we are left with circularity. Another way of putting this is to say that if the set of facts which determine, say, moral responsibility attributions do not include facts about moral responsibility attributions, then there is no circle of dependency. So, to expose any circularity we should select a determinant, lay out what kind of facts are its determiners, and identify any overlap between the two. Let's start with facts about moral responsibility.

In Chapter 4, I argued that whether one is morally responsible for an outcome through a failure to act depends upon two kinds of causal facts: facts about causal stability and facts of the kind '*c* caused *e*'. I argued that the kind of causal stability which determines moral responsibility attributions does not appeal to moral facts. Hence, we can set aside this branch as a potential avenue for circularity. What about the other branch? In Chapters 2 and 3, I argued that facts of the kind '*c* caused *e*' are determined by the concept of normality, and further that the concept of normality is comprised of (amongst other things) moral norms. Here then is where circularity could creep in. If the kind of moral norms that determine normality include facts about moral responsibility, then facts about moral responsibility will end up being determined by facts about moral responsibility. Thus arises circularity.

However, there's no overlap here: the kind of moral norms we're interested in when determining facts like '*c* caused *e*' are not norms pertaining to moral responsibility. Take the canonical moral norm used in the causation literature — promise keeping. It is thought that whether your friend promises to feed your fish makes a difference as to whether they are a cause of the fish's death when it eventually dies from a lack of food, because causal relations are sensitive to norm violations and breaking a promise constitutes a violation of a moral norm. These types of violations do not necessarily amount to judgements of moral responsibility. Your friend could break her promise by failing to feed the fish, but if the reasons she failed to feed them are no fault of hers then surely she cannot be held morally responsible for the promise breaking. Suppose she didn't feed them because you had mistakenly told her to do so in late

[72] By 'determinant' I mean the thing to be determined. We can think of the determiner/determinant relationship as similar to the explanandum/explanans relationship, but I stick with the terminology of determinacy to avoid confusion.

March when you're on holiday, but you actually go on holiday in early March and on your return you're met with dead fish. Although your friend violates a moral norm insofar as she fails to keep her promise, she cannot be held morally responsible for the consequences of that failure given the reasons for which she broke it.

To illustrate this idea further consider a different moral norm invoked to determine the truth-value of '*c* caused *e*'. This time consider the kind of moral norms invoked by Knobe and Fraser (2008) in their well-known pen vignette case. To recap the case: the receptionist in the philosophy department keeps her desk stocked with pens, the admin staff are allowed to take the pens but faculty members are supposed to buy their own. One morning Professor Smith is walking past the desk and takes the last pen. Later that day, the receptionist needs to take an important message but cannot because she has no pens left. Intuitively Professor Smith caused the problem, and the reasons she is a cause (as opposed to the admin staff) is because her pen-taking violates a moral norm: faculty are not allowed to take pens. So, does Professor Smith's norm violation necessarily amount to a judgement about her moral responsibility? No. Imagine that Professor Smith is a new member of faculty, she is thus not aware of the department's rules around pen distribution and it would be unreasonable to expect her know of such rules. In this story, Professor Smith is presumably not morally responsible for the problem even though she in fact violates a moral norm. If this is the right way to characterise things, then the kind of moral norms being appealed to in order to determine facts of the type '*c* caused *e*' do not include facts about moral responsibility. As a result, facts about moral responsibility are not determined by facts about moral responsibility, meaning there is no circularity of dependency here.

Next, let's check for circularity in the other set of moral facts explored in the thesis — facts about moral worth. In Chapter 6, I argued that moral worth depended upon how causally robust one's motive was for acting, and I argued that causal robustness depends upon how many normal worlds the agent's motive would produce right action in. I characterised a normal world in roughly the same way I characterised the notion of normality in the first two Chapters, i.e. comprised of various norms including moral norms. Here again is where circularity could creep in. If the moral norms that make a world normal include considerations about moral praiseworthiness, then facts about moral praiseworthiness will be determined by facts about moral praiseworthiness.

It seems quite straightforward to me that norms about praiseworthiness (if indeed there are such norms) are not the norms that make a world normal. A world is not made morally normal in virtue of an agent being praiseworthy. To illustrate, we can suppose that a world in which people keep their promises is more normal, other things being equal, than a world where people break them. But this promise-keeping-world isn't made more normal if agents are also praiseworthy for keeping their promises. Adhering to moral norms doesn't require that the agent be motivated in a certain way that makes them praiseworthy.[73]

To summarise then, this thesis has endeavoured to establish the interdependency between causation and morality. It has argued that certain kinds of causal facts partly determine certain kinds of moral facts, and that certain kinds of moral facts partly determine certain kinds of causal facts. But given that the facts being determined are not the same as the determiners, this interdependency does not entail a circularity.

[73] In demonstrating that there is no circularity running from moral facts to causal facts, I have also established that there is no circularity running the other way from causal facts to moral facts given the very nature of circularity.

Bibliography

Arpaly, N and Schroeder, T. (2013). *In Praise of Desire*. Oxford: Oxford University Press.

Arpaly, N. (2002). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Beebee, H. (2004). Causing and Nothingness. In L. A. Paul, N. Hall, & J. Collins (Eds.), *Causation and Counterfactuals* (pp. 291–308). MIT Press.

Beebee, H. and Kaiserman, A. (2020), Causal Contribution in War. *Applied Philosophy*, 37: 364-377. https://doi.org/10.1111/japp.12341

Bennett, J. (1974). The Conscience of Huckleberry Finn. *Philosophy*, 49(188), 123–134.

Benson, P. (1987). Moral worth. *Philosophical Studies*, 51(3), 365–382.

Bernstein, S. (2014). Omissions as possibilities. *Philosophical Studies* 167 (1):1-23.

Bernstein, S. (2016). Causal and Moral Indeterminacy. *Ratio* 29(4):434-447.

Bernstein, S. (2017a). Causal Idealism. In T. Goldschmidt & K. L. Pearce (Eds.), *Idealism: New Essays in Metaphysics*. Oxford University Press. https://doi.org/10.1093/oso/9780198746973.001.0001

Bernstein, S. (2017b). Causal Proportions and Moral Responsibility. In David Shoemaker (ed.), *Oxford Studies in Agency and Responsibility, Volume 4*. pp. 165-182. Oxford: Oxford University Press.

Blanchard, T. & Schaffer, J. (2017). Cause without Default. In Helen Beebee, Christopher Hitchcock & Huw Price (eds.), *Making a Difference*. pp. 175-214.Oxford: Oxford University Press.

Braham, M., & Van Hees, M. (2012). An Anatomy of Moral Responsibility. *Mind*, *121*(483), 601–634. https://doi.org/10.1093/mind/fzs081

Braham, M., van Hees, M. (2009) Degrees of Causation. *Erkenntnis* 71, 323–344

Byrd, J. (2007). Moral Responsibility and Omissions. *The Philosophical Quarterly*, 57(226), 56–67.

Catita, M., Águas, A. & Morgado, P. (2020) Normality in medicine: a critical review. *Philosophy Ethics Humanities Medicine* 15, 3. https://doi.org/10.1186/s13010-020-00087-2

Campbell, J. (2007). An interventionist approach to causation in psychology. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 58–66). Oxford University Press.

Chadwick, R. (2016). Normality as Convention and as Scientific Fact. In Schramme, T and Edwards, S. (eds) *Handbook of Philosophy of Medicine*. Springer https://doi.org/10.1007/978-94-017-8706-2_9-1

Clarke, R. (1994). Ability and Responsibility for Omissions. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *73*(2/3), 195–208.

Clarke, R. (2014). *Omissions*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199347520.003.0002

Demirtas H. (2022) Causation comes in degrees. *Synthese*. 200(1):1-17

Dowe, P. (2000). *Physical Causation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511570650

Driver, J. (2008). Attributions of causation and moral responsibility. In *Moral psychology, Vol 2: The cognitive science of morality: Intuition and diversity* (pp. 423–439). MIT Press.

Eells, E. (1991). *Probabilistic Causality*. Cambridge University Press. https://doi.org/10.1017/CBO9780511570667

Feinberg, J. (1970). *Doing and deserving: Essays in the theory of responsibility*. Princeton University Press.

Feinberg, J. (1987). *The Moral Limits of the Criminal Law Volume 1: Harm to Others*. Oxford University Press. https://doi.org/10.1093/0195046641.001.0001.002.012

Ferguson, B. (2012). Kant on Duty in the Groundwork. *Res Publica*, 18, 303–319.

Franklin-Hall, L. R. (2016). High-Level Explanation and the Interventionist's 'Variables Problem'. *British Journal for the Philosophy of Science* 67 (2):553-577.

Frisch, M. (2014). *Causal Reasoning in Physics*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139381772

Garvey, J. (2011). *Climate Change and Causal Inefficacy: Why Go Green When It Makes No Difference?* Royal Institute of Philosophy Supplement, 69, 157-174. doi:10.1017/S1358246111000269

Glennan, S. (2002). Rethinking Mechanistic Explanation. *Philosophy of Science*, 69(S3), 342–353. https://doi.org/10.1086/341857

Grinfeld ,G. Lagnado, D. Gerstenberg, T. Woodward J, F. and Usher, M. (2020). Causal Responsibility and Robust Causation. *Frontiers of Psychology*. 11(1069). doi:10.3389/fpsyg.2020.01069

Hall, N. (2006). Comments on Woodward, "Making Things Happen" [Review of *Making Things Happen: a Theory of Causal Explanation*, by J. Woodward]. *History and Philosophy of the Life Sciences*, *28*(4), 611–624.

Hall, N. (2007). Structural Equations and Causation. *Philosophical Studies,* 132(1), 109–136.

Hall, N., & Paul, L. A. (2013). Metaphysically Reductive Causation. *Erkenntnis*, *78*(S1), 9–41. https://doi.org/10.1007/s10670-013-9435-6

Hall, Ned & Paul, Laurie Ann (2003). Causation and preemption. In Peter Clark & Katherine Hawley (eds.), *Philosophy of Science Today*. Oxford University Press.

Halpern, J. Y. (2016). *Actual Causality*. The MIT Press. https://doi.org/10.7551/mitpress/9780262035026.001.0001

Halpern, J. Y., & Hitchcock, C. (2010). Actual causation and the art of modelling. In H. Geffner, J. Halpern Y., & R. Dechter (Eds.), *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (pp. 383–406). College Publications

Halpern, J. Y., & Hitchcock, C. (2015). Graded Causation and Defaults. *The British Journal for the Philosophy of Science*, 66(2), 413–457. https://doi.org/10.1093/bjps/axt050

Hart, H. L. A., & Honoré, T. (1985). *Causation in the Law*. Second Edition. Oxford University Press.

Henson, R.G. (1979). What Kant Might Have Said: Moral Worth and the Overdetermination of Dutiful Action. *The Philosophical Review*, 88(1), 39-54.

Herman, B. (1981). On the Value of Acting from the Motive of Duty. *The Philosophical Review*, 90(3), 359-382.

Hitchcock, C. (2007). Prevention, Preemption, and the Principle of Sufficient Reason. *The Philosophical Review*, *116*(4), 495–532.

Hitchcock, C. and Knobe, J. (2009) Cause and Norm. *Journal of Philosophy*, 11, 587-612.

Hitchcock, C. and Woodward, J. (2003) 'Exploratory Generalization Part II: Plumbing Exploratory Depth', *Nôus*, 37: 181–99

Honoré, A. M. (1999). *Responsibility and Fault*. Bloomsbury Publishing.

Hume, D. (1969). *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*. (Ed) Mossner, E. Penguin Group

Husak, D. (2017). Courses of Conduct. In *The Ethics and Law of Omissions*. Oxford University Press. https://doi.org/10.1093/oso/9780190683450.003.0009

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, 161, 80–93. https://doi.org/10.1016/j.cognition.2017.01.010

Isserow, J. (2019). Moral Worth and Doing the Right Thing by Accident. *Australasian Journal of Philosophy*, 97(2), 251-264.

Kaiserman, A. (2016). Causal Contribution. *Proceedings of the Aristotelian Society* 116 (3):387-394.

Kaiserman, A. (2018). 'More of a Cause': Recent Work on Degrees of Causation and Responsibility. *Philosophy Compass* 13 (7):12498.

Kant, I. (1997). *Groundwork of the Metaphysics of Morals*. (trans.)(ed.) Gregor, M. United Kingdom: Cambridge University Press.

Kim, Jaegwon (1984). Concepts of supervenience. *Philosophy and Phenomenological Research*, 45,153-76.

Kim, J. (1988). Explanatory Realism, Causal Realism, and Explanatory Exclusion. *Midwest Studies In Philosophy*, 12(1), 225–239. https://doi.org/10.1111/j.1475-4975.1988.tb00167.x

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 441–447). MIT Press.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137, 196–209. https://doi.org/10.1016/j.cognition.2015.01.013

Kutz, C. (2000) *Complicity: Ethics and Law for a Collective Age*. Cambridge: Cambridge University Press

Kutz, C. (2007). Causeless complicity. *Criminal Law and Philosophy* 1(3):289-305.

Leuridan, Bert (2010) 'Can Mechanisms Really Replace Laws of Nature?',

Lewis, D. (1973). Causation. *The Journal of Philosophy*, 70(17), 556–567. https://doi.org/10.2307/2025310

Lewis, D. (1986). Postscript C to "Causation": Insensitive causation. Philosophical Papers. Vol. 2 (pp. 184–188). Oxford, UK: Oxford University Press.

Lewis, D. (1987). Causation. In *Philosophical Papers Volume II*. Oxford University Press. https://doi.org/10.1093/0195036468.003.0006

Lewis, D. (2000). Causation as influence. *Journal of Philosophy* 97(4):182-197.

Lippert-Rasmussen, K. (2021). Praising Without Standing. *The Journal of Ethics* 26(2):229-246

Lord, E. (2017). What You're Rationally Required to Do and What You Ought to Do (Are the Same Thing!). *Mind*, 126(504),1109–1154.

Mackie, J. L. (1980). *The Cement of the Universe: A Study of Causation*. Oxford University Press. https://doi.org/10.1093/0198246420.001.0001

Markovits, J. (2010). Acting for the Right Reasons. *Philosophical Review*, 119(2), pp.201–242.

Markovits, J. (2012). Saints, Heroes, Sages, and Villains. *Philosophical Studies*, 158(2), 289–311.

Massoud, A. (2016). Moral Worth and Supererogation. *Ethics*, 126(3), 690–710.

McDonnell, N. (2019). The Non-Occurrence of Events. *Philosophy and Phenomenological Research*, 99(2), 269–285. https://doi.org/10.1111/phpr.12476

McGrath, S. (2005). Causation By Omission: A Dilemma. *Philosophical Studies*, 123(1–2), 125–148. https://doi.org/10.1007/s11098-004-5216-z

Menzies, P. (2004). Difference Making in Context. In J. Collins, N. Hall, & P. Laurie (Eds.), *Causation and Counterfactuals*. MIT Press.

Menzies, P. (2006). Review: Making Things Happen: A Theory of Causal Explanation. *Mind*, 115(459), 821–826. https://doi.org/10.1093/mind/fzl821

Menzies, P. (2007). Causation in Context. In R. Corry & H. Price (Eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press.

Menzies, P. (2009). Platitudes and Counterexamples. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *The Oxford Handbook of Causation* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199279739.003.0018

Metz, J. (2021). An ability-based theory of responsibility for collective omissions. *Philosophical Studies*, 178(8), 2665–2685. https://doi.org/10.1007/s11098-020-01568-y

Miller, D.J. (2018). Circumstantial Ignorance and Mitigated Blameworthiness. *Philosophical Explorations*, 22(1), 33–43.

Montmarquet, J. (2012). Huck Finn, Aristotle, and Anti-Intellectualism in Moral Psychology. *Philosophy*, 87(1), 51–63.

Moore, M, S. (2009). *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics.* Oxford Press

Murray, D. and Lombrozo, T. (2017), Effects of Manipulation on Attributions of Causation, Free Will, and Moral Responsibility. *Cognitive Science*, 41: 447-481. https://doi.org/10.1111/cogs.12338

Nelkin, D. K., & Rickless, S. C. (2017). *The Ethics and Law of Omissions*. Oxford University Press. https://doi.org/10.1093/oso/9780190683450.001.0001

Northcott, R. (2013) Degree of explanation. *Synthese* 190, 3087–3105

Oshana, M. A. L. (1997). Ascriptions of Responsibility. *American Philosophical Quarterly*, 34(1), 71–83.

Paul, L. A. (2000). Aspect Causation. *The Journal of Philosophy*, 97(4), 235–256. https://doi.org/10.2307/2678392

Pearl, J. (2000) *Causality*. Cambridge University Press

Pereboom, D. (2015). Omissions and Different Senses of Responsibility. In A. Buckareff, C. Moya, & S. Rosell (Eds.), *Agency, Freedom, and Moral Responsibility* (pp. 179–191).

Palgrave Macmillan UK. https://doi.org/10.1057/9781137414953_12

Petersson, B. (2013). Co-responsibility and Causal Involvement. *Philosophia* 41(3):847-866.

Prescott-Couch, Alexander (2017). Explanation and Manipulation. *Noûs* 51 (3):484-520.

Putnam, H. (1982). Why There Isn't a Ready-Made World. *Synthese*, *51*(2), 141–167.

Reutlinger, A. (2013). *A Theory of Causation in the Social and Biological Sciences*. Palgrave Macmillan UK. https://doi.org/10.1057/9781137281043

Rosen, G. (2004). Skepticism about Moral Responsibility. *Philosophical Perspectives*, 18(1), 295–313. https://doi.org/10.1111/j.1520-8583.2004.00030.x

Russell, B. (1912). On the Notion of Cause. *Proceedings of the Aristotelian Society*, *13*, 1–26.

Sartorio, C. (2004). How To Be Responsible For Something Without Causing It. *Philosophical Perspectives*, 18(1), 315–336. https://doi.org/10.1111/j.1520-8583.2004.00031.x

Sartorio, C. (2007). Causation and Responsibility. *Philosophy Compass*, *2*(5), 749–765. https://doi.org/10.1111/j.1747-9991.2007.00097.x

Sartorio, C. (2016). *Causation and Free Will*. Oxford University Press UK.

Sartorio, C. (2019), More of a Cause?. *Applied Philosophy*, 37: 346-363. https://doi.org/10.1111/japp.12370

Scanlon, T. (2008). *Moral dimensions: Permissibility, meaning, blame*. Harvard University Press.

Schaffer, J. (2004). Causes need not be Physically Connected to their Effects: The Case for Negative Causation. In C. Hitchcock (Ed.), *Contemporary Debates in the Philosophy of Science*. Blackwell Publishing.

Schaffer, J. (2010). Contrastive causation in the law. *Legal Theory* 16 (4):259-297.

Shafer-Landau, R. (2003). *Moral Realism: A Defence*. Oxford University Press. https://doi.org/10.1093/0199259755.001.0001

Shapiro, L., & Sober, E. (2007). Epiphenomenalism–The Do's and Don'ts. In G. Wolters & P. Machamer (Eds.), *Studies in causality: historical and contemporary* (pp. 235–264). Pittsburgh: University of Pittsburgh Press.

Shoemaker, D. (2015). Answerability. In D. Shoemaker (Ed.), *Responsibility from the Margins* (p. 0). Oxford University Press.

Sinnott-Armstrong, W. (2005). It's Not My Fault: Global Warming and Individual Moral Obligations. In Walter Sinnott-Armstrong & Richard Howarth (eds.), *Perspectives on Climate Change*. (pp. 221–253) .Elsevier.

Sliwa, P. (2016). Moral Worth and Moral Knowledge. *Philosophy and Phenomenological Research*, 93(2), 393–418.

Smith, A. M. (2015). Responsibility as Answerability. *Inquiry*, *58*(2), 99–126. https://doi.org/10.1080/0020174X.2015.986851

Smith, A. M. (2017). Unconscious Omissions, Reasonable Expectations, and Responsibility, in Dana Kay Nelkin, and Samuel C. Rickless (eds), *The Ethics and Law of Omissions* https://doi.org/10.1093/oso/9780190683450.003.0003

Smith, M. (2007). Ceteris Paribus Conditionals and Comparative Normalcy. *Journal of Philosophy of Logic* 36, 97-121.

Smith, M. (2010). What Else Justification Could Be1. *Noûs*, 44(1), 10-31.

Smith, P. G. (1990). Contemplating Failure: The Importance of Unconscious Omission. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *59*(2), 159–176.

Sorensen, K. (2010). Effort and Moral Worth. *Ethical Theory and Moral Practice*, 13(1), 89–109.

Statham, G. (2020). Causes as Deviations from the Normal: Recent Advances in the Philosophy of Causation. In E. A. Bar-Asher Siegal & N. Boneh (Eds.), *Perspectives on Causation: Selected Papers from the Jerusalem 2017 Workshop* (pp. 445–462). Springer International Publishing.

Stratton-Lake, P. (2000). *Kant, Duty and Moral Worth*. London: Routledge.

Varzi, A. (2007). Omissions and Causal Explanations. In Francesca Castellani & Josef Quitterer (eds.), *Agency and Causation in the Human Sciences*. (pp. 155–167). Mentis Verlag.

Vasilyeva, N., Blanchard, T. and Lombrozo, T. (2018), Stable Causal Relationships Are Better Causal Relationships. *Cognitive Science*, 42: 1265-1296.

Walton, D. (1980). Omitting, Refraining and Letting Happen. *American Philosophical Quarterly*, 17, 319–326.

Waters, C. K. (2007). Causes That Make a Difference. *Journal of Philosophy* 104 (11):551-579. DOI: 10.5840/jphil2007104111

Whittle, A. (2018). Responsibility in Context. *Erkenntnis*, 83(2), 163–183. https://doi.org/10.1007/s10670-017-9884-4

Wolf, S. (2015). Responsibility, Moral and Otherwise. *Inquiry*, 58(2), 127–142. https://doi.org/10.1080/0020174X.2015.986852

Woodward, J (2001). *'Causation and Manipulability'*. (n.d.). Retrieved 13 October 2022, from https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=causation-mani

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, J. (2006). Sensitive and Insensitive Causation. *The Philosophical Review*, 115(1), 1–50. https://doi.org/10.1215/00318108-2005-001

Woodward, J. (2014). A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters—Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment). *Philosophy of Science*, 81(5), 691-713. doi:10.1086/678313

Woodward, J. (2016). The problem of variable choice. *Synthese*, *193*(4), 1047–1072. https://doi.org/10.1007/s11229-015-0810-5