Actual Causation

Von der Philosophischen Fakultät
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades
Doktor der Philosophie (Dr. phil.) genehmigte
Dissertation

von Enno Fischer

Erscheinungsjahr: 2021

Referent: Prof. Dr. Mathias Frisch, Institut für Philosophie, Leibniz Universität Hannover

Korreferent: Prof. Dr. Christopher Hitchcock, Division of the Humanities and Social Sciences, California Institute of Technology

Tag der Promotion: 03.08.2020

Abstract

In this dissertation I develop a pluralist theory of actual causation. I argue that we need to distinguish between total, path-changing, and contributing actual causation. The pluralist theory accounts for a set of example cases that have raised problems for extant unified theories and it is supported by considerations about the various functions of causal concepts. The dissertation also analyses the context-sensitivity of actual causation. I show that principled accounts of causal reasoning in legal inquiry face limitations and I argue that the context-sensitivity of actual causation is best represented by a distinction between default and deviant states in causal models.

The dissertation has three parts. Part I provides the theoretical background for my pluralist account. First, two central problems for theories of actual causation are introduced: the problem of redundancy and the problem of selection (Chapter 2). Then I review extant accounts that address these problems by employing the framework of causal models. A central assumption of these accounts is that there is a single unified concept of actual causation (Chapter 3).

Part II presents my pluralist account. I argue from an interventionist perspective that we need to distinguish between total, path-changing, and contributing actual causation. Moreover, I present a novel distinction between two senses in which concepts of actual causation are context sensitive. Context-sensitivity$_1$ concerns the normality or typicality of an individual event, independently of other events that occur. Context-sensitivity$_2$ concerns our willingness to consider an event's occurrence given that certain other events also occur (Chapter 4). Next, I argue that we also need to be pluralist with regard to the function of concepts of actual causation. I show that interventionist approaches to the function of causal concepts face limitations. In order to make sense of certain claims of actual causation we need to assume that their function is to indicate responsibility for some outcome (Chapter 5).

Part III explores consequences of the pluralist account with particular regard to the context-sensitivity of actual causation. First, I employ the distinction between two kinds of context-sensitivity, in order to show that principled approaches to causal inquiry in the law face difficulties. While principled approaches may circumvent context-sensitivity$_1$, it is unlikely that they will overcome context-sensitivity$_2$ (Chapter 6). Finally, I present a new argument for incorporating a distinction between default and deviant states into the formal framework of causal models. Sometimes agents agree about the causal structure but disagree about the appropriate causal judgments. In this kind of situations causal models with defaults can facilitate clarification by enabling a distinction between epistemic disagreement about causal structure and normative disagreement about relevant possibilities (Chapter 7).

Keywords: causal models, causal pluralism, interventionist accounts of causation, responsibility, context-sensitivity.

Zusammenfassung

In dieser Dissertation entwickle ich eine pluralistische Theorie der Verursachung. Ich argumentiere für eine Unterscheidung zwischen totaler, pfad-verändernder und beitragender Verursachung. Die pluralistische Theorie löst Beispielfälle, die für existierende vereinheitlichende Theorien der Verursachung Schwierigkeiten hervorrufen. Außerdem wird die Theorie durch Überlegungen zu den vielseitigen Funktionen kausaler Begriffe untermauert. Ferner analysiert die Dissertation die Kontextabhängigkeit von Verursachung. Ich weise Grenzen für prinzipienbasierte Ansätze zur rechtlichen Ursachenanalyse auf und ich lege ein neues Argument dafür vor, dass kausale Modelle kontextabhängige Überlegungen am besten mit Hilfe einer Unterscheidung zwischen Standardzuständen und abweichenden Zuständen repräsentieren.

Die Dissertation hat drei Teile. Teil I legt den theoretischen Hintergrund für meinen pluralistischen Ansatz dar. Zwei zentrale Probleme für Theorien der Verursachung werden eingeführt: das Problem der Redundanz und das Problem der Auswahl (Kapitel 2). Anschließend gebe ich einen Überblick über existierende Ansätze, die diese Probleme mit Hilfe kausaler Modelle zu lösen versuchen. Eine zentrale Annahme dieser Ansätze ist, dass es einen einzigen vereinheitlichten Begriff von Verursachung gibt (Kapitel 3).

Teil II präsentiert meinen pluralistischen Ansatz. Aus einer interventionistischen Perspektive argumentiere ich dafür, dass wir zwischen totaler, pfad-verändernder und beitragender Verursachung unterscheiden müssen. Ich führe außerdem eine neue Unterscheidung zwischen zwei Arten von Kontextabhängigkeit von Verursachung ein. Kontextabhängigkeit$_1$ betrifft die Normalität oder Typikalität eines individuellen Ereignisses, unabhängig von anderen Ereignissen. Kontextabhängigkeit$_2$ betrifft unsere Bereitschaft, das Eintreten eines bestimmten Ereig-nisses zu erwägen, gegeben, dass bestimmte andere Ereignisse eintreten (Kapitel 4). Danach argumentiere ich dafür, dass wir auch eine pluralistische Theorie für die Funktion von Verursachungsbegriffen brauchen. Ich zeige, dass interventionistische Ansätze zur Funktion von Verursachungsbegriffen eine begrenzte Erklärungskraft haben. Um bestimmte Aussagen über Verursachung richtig einzuordnen, müssen wir annehmen, dass sie die Funktion haben, auf Verantwortungsträger hinzuweisen.

Teil III lotet die Konsequenzen meines pluralistischen Ansatzes aus und widmet sich der Kontextabhängigkeit von Verursachung. Zuerst verwende ich die Unterscheidung zwischen zwei Arten von Kontextabhängigkeit, um zu zeigen, dass prinzipienbasierte Ansätze zur rechtlichen Ursachenanalyse mit Problemen konfrontiert sind. Während prinzipienbasierte Ansätze möglicherweise Kontextsensitivität$_1$ umgehen können, ist es unwahrscheinlich, dass sie Kontextabhängigket$_2$ verhindern können (Kapitel 6). Schließlich präsentiere ich ein neues Argument für eine Unterscheidung zwischen Standardzuständen und abweichenden Zuständen in kausalen Modellen. Manchmal sind sich Akteure über die kausale Struktur einig, während sie sich uneinig über die angemessenen Kausalurteile sind. In diesen Situationen können kausale Modelle, die Standardzustände und abweichende Zustände unterscheiden, Klarheit verschaffen. Sie ermöglichen eine Unterscheidung zwischen epistemischer Uneinigkeit über die kausale Struktur von normativer Uneinigkeit über relevante Möglichkeiten.

Schlagwörter: Kausale Modelle, kausaler Pluralismus, interventionistische Theorien der Kausalität, Verantwortung, Kontextabhängigkeit.

Acknowledgments

First of all, I would like to thank Mathias Frisch and Hasok Chang for giving me the opportunity to conduct this project under their supervision. Thank you very much for your feedback, advice, and support.

I am particularly grateful to Christopher Hitchcock for acting as external referee for the thesis. As the following pages show, his work has had an enormous influence on my thinking about actual causation.

My intellectual home during the past years has been the DFG Graduiertenkolleg 2073 "Integrating Ethics and Epistemology of Scientific Research." One of the great advantages of the Graduiertenkolleg has been the opportunity to benefit from a large group of members of faculty both at the University of Hanover and the University of Bielefeld. I would like to thank all principle investigators for useful comments on my presentations at the GRK colloquia and many helpful conversations. In particular, I would like to thank Uljana Feest and Marie Kaiser for detailed comments on drafts of chapters of the thesis and for helpful advice during tutorials.

Life as a PhD student would surely have been much poorer had it not been for the wonderful community of fellow PhD students and Postdocs within the GRK. I am grateful for joint enjoyment of the culinary highlights provided by the Contine, long evenings of table tennis at the theory workshops, and an open ear for my philosophical concerns even after the 1000th iteration of the notorious thought experiments involving Billy and Suzy. Many thanks especially to the

Hanover-based part of the group, including Markus Ahlers, Stefano Canali, Philipp Eichenberger, David Hopf, Daria Jadreškić, Anna Leuschner, Jonas Lipski, Leon Schäfer, Julia Pfeiff, and Tobias Schönwitz. Many thanks also to Rafael Ugarte Chacón for being so extremely helpful in navigating any administrative issue that I encountered during the time of my PhD.

Being part of the Institute of Philosophy at the Leibniz Universität Hannover has been a wonderful experience. I am so grateful for having benefited from the open intellectual atmosphere and the many helpful conversations with members of the Institute of Philosophy. Many thanks to Philippe van Baßhuysen, Radin Dardashti, Donal Khosrowi, Sebastian Krug, Reuben Stern, and Lucie White for illuminating discussions and/or feedback on drafts.

I had the great privilege of being able to spend Michaelmas Term 2019 at the Department of History and Philosophy of Science at the University of Cambridge. Many thanks to Hasok Chang for making this stay possible and so enjoyable. I would also like to thank Jeremy Butterfield for helpful advice and wonderful conversations during this time. Special thanks to Grace Field, Céline Henne, Sarah Hijmans, Paul Sampson, Emilie Skulberg, Hannah Tomczyk, Bobby Vos, and Adrian Yee for making my research stay such an intellectually stimulating and fun experience. I would also like to thank the DAAD for supporting the research stay at the University of Cambridge with a generous grant.

I am also grateful for having been able to visit the Munich Center for Mathematical Philosophy for a couple of times in the beginning of 2018. Many thanks to Stephan Hartmann for making these research visits possible and many thanks to Alexander Reutlinger, Reuben Stern, and Michele Luchetti for helpful exchange.

I consider the workshop on "Causes, Norms, and Decisions" to be one of the academic highlights of my time as a PhD student. Many thanks to Mathias Frisch for making this workshop possible and many thanks to Sebastian Krug for coorganizing the workshop. Special thanks for helpful conversations and feedback to Jon Bebb, Sander Beckers, Thomas Blanchard, Alison Fernandes, Christopher Hitchcock, Weyma Lübbe, Huw Price, and Georgie Statham.

Many thanks to Sören Hilbrich and Tammo Lossau for countless philosophical discussions and for feedback on the thoughts and drafts leading up to this thesis.

I am grateful to my parents Rudolf Fischer and Ilona Kändler-Fischer who have always supported and encouraged me on the path that has led to writing this thesis. I also want to thank my grandmother Edeltrud Kändler who was always interested in this work and who would have been so glad to see the result.

Managing the everyday life of a young family can sometimes be a challenge, especially in times of hard work. Many thanks to my parents in law Lyubov Kurdakova and Valeriy Kurdakov for their wonderful help.

Finally, I am deeply grateful to Anastasiia Kurdakova for her love, her unconditional support, and for reminding me that there are other things in life besides philosophy.

Contents

1. Introduction

The topic of this thesis is the concept of actual causation. The term 'actual' suggests that this concept refers to a particular kind of cause—a kind of cause that is naturally contrasted with potential causes. A potential cause is something that *can* bring about a certain effect. An actual cause then is something that *does* bring about a certain effect.

The distinction between actual and potential causes is relevant in situations where we want to know why a certain event occurred or a certain state of affairs came about. Often we are able to identify a number of conditions that could explain the outcome but we need to know what actually did bring it about. Consider, for example, a legal inquiry that concerns a particular person's death. Common sense suggests a wide range of potential causes. For example, given the circumstantial evidence it may be hypothesized that the person's death was caused by poisoning. In order to verify the hypothesis, we need to know not only that poisoning is one way in which a person can die. We need to know whether the person's death was actually caused by poison.

Here is another example—one that suggests that actual causation is relevant in contexts of intervention as well. Suppose you have serious back pain and visit your doctor. The doctor knows about a wide range of potential causes of back pain: poor posture, little exercise, or more serious conditions such as a prolapsed disk or a broken bone in the spine. In order to prescribe the optimal medical intervention, the doctor will need to know which of these factors actually causes your back pain.

Actual causes matter in the sciences as well. In the 1960s Arno Penzias and Robert Woodrow Wilson, two radio astronomers at Bell labs, measured an excess signal of 2.7 K when they directed their horn antenna at the sky (Penzias and Wilson (1965)). They had a range of hypotheses about potential causes of the deviation, including human-made radiation from New York City, ionized particles in the atmosphere resulting from high altitude atomic bomb tests, and pigeon's droppings in the antenna (Wilson (1978)). However, they wanted to know what *did* cause the deviation. Eventually they concluded that the antenna was picking up fossil radiation from an early stage of the universe, today known as the Cosmic Microwave Background Radiation.

What distinguishes actual causes from merely potential causes? Actual causes do occur. The prolapsed disk is an actual cause of your back pain only if it is in fact prolapsed. Likewise, the poisoning is an actual cause of the victim's death only if the person was really poisoned and the Microwave Background is an actual cause only if it is true that there is background radiation of this type.

But mere realization of the respective events or states of affairs is not sufficient. The cause also has to bring about the effect. A natural way to spell out what this means is to say that the effect depends on the cause. If the cause had not occurred, then the effect would not have occurred either. However, this criterion is known to fail in instances where more than one factor could have brought about the effect. Consider, for example, the following case of preemption. An assassin in training shoots the victim. On his mission he is accompanied by a supervising assassin who would have shot the victim if he had lost his nerves. This is a situation where the effect (the victim's death) does not depend on its cause (the trainee's pulling the trigger of his gun) because there is a backup process (the supervisor) that would have sustained the effect if the primary process had failed (Hitchcock, 2001).

One approach to cases like these, going back to Judea Pearl (2000), is that counterfactual dependence as a criterion for causation can be retained if we require dependence *under certain contingencies*. Under the contingency that the supervising assassin does not intervene, the victim's death depends on the trainee's behaviour, therefore the trainee is an actual cause. The crucial challenge for an account like this is to find general criteria that delineate permissible from impermissible contingencies. There is now a wide range of accounts that suggest different criteria for the delineation. Most accounts of this kind spell out the criteria with the help of causal models. A causal model consists of a set of variables and a set of structural equations. The variables represent possible events or states of affairs in the target system and the structural equations describe the causal relations between the variables. Joseph Halpern's book *Actual Causality* (2016) is exemplary for the current state of art in the causal-model based literature on actual causation. Surveying a plethora of example cases Halpern offers three competing basic definitions of actual causation and discusses several ways of combining these definitions with additional criteria that restrict contingencies in terms of "normality." At the same time, Halpern is convinced that there is one unified concept of actual causation that captures the causal intuitions in the wealth of example cases.

In this thesis I will propose a new account of actual causation that is explicitly pluralist. According to my account, there is a range of concepts of actual causation that need to be distinguished. More specifically, I shall propose a distinction between three notions of actual causation. In order to provide a preliminary understanding of the distinction let me give a brief and rough characterization of the concepts—exact definitions will be provided in Chapter 4. First, an event c is a *total actual cause* of another event e if and only if e depends on c. This means that if c is a total actual cause of e, an intervention that prevents c is sufficient in order to prevent e.

Second, an event c is a *path-changing actual cause*[1] of another event e if and only if e depends on c, given that certain other consequences of intervening on c do not occur. The key difference to total actual causation is that in order to prevent the effect of a path-changing actual cause we need to apply (at least) two interventions. If c is a path-changing actual cause of e, then a primary intervention that prevents c needs to be combined with a secondary intervention on adverse consequences of the intervention on c. For example, an intervention on the assassin in training would lead to an attack on the victim by the supervising assassin. In order to save the victim the primary intervention on the assassin in training has to be combined with a secondary intervention on the supervising assassin.

Third, an event c is a *contributing actual cause*[2] of another event e if and only if e depends on c, given that certain other events c' that are independent of c do not occur. This means that if c is a contributing actual cause of e (but not a path-changing or total actual cause), then an intervention that prevents c has to be combined with an intervention that prevents c' in order to prevent e. Contributing actual causation is found in examples of symmetrical overdetermination. Each one of two fires that approach a house from different sides is a contributing actual cause of the house being destroyed. Like in path-changing actual causation the goal (e.g. saving the house) is achieved only if more than one intervention is applied. The difference is that the additional intervention targets a causal process that is independent of the process targeted by the first intervention and was active in the first place.[3] This is relevant because such active causal processes are often much easier to track than processes that are only activated through the attempt of preventing the effect.

[1] The definition of path-changing actual cause that I will propose in Chapter 4 will be similar to (and is inspired by) definitions of actual cause provided by Pearl (2000), Hitchcock (2001), and Halpern (2015).

[2] The definition of contributing actual cause that I will propose in Chapter 4 will be similar to (and is inspired by) Halpern and Pearl's (2005) preliminary definition of actual cause.

[3] More precisely, this is true for instances of contributing actual causation that are not instances of path-changing actual causation.

Causal pluralism is typically motivated through failures to capture the diversity of causal reasoning within one unified account of causation. Nancy Cartwright (2007), for example, argues that there is a range of opposing approaches to causation which solve certain paradigmatic problem cases but fail with respect to others. This observation is taken to support the pessimistic inference that no account will ever achieve unification. My pluralist theory is not based on such an assumption.

Instead, I propose, we need to distinguish a plurality of causal concepts in order to facilitate the functions of causal reasoning most effectively. In particular, I argue that the differences between total, path-changing, and contributing actual causation are important from the perspective of an intervening agent. If we suppose that actual causation has the function to indicate intervention targets that are most suited for achieving our goals (as argued, for example, by Christopher Hitchcock and Joshua Knobe (2009)), then the concept should *at least* tell us whether a single intervention is sufficient for achieving the goal or whether that intervention needs to be combined with another intervention.

Thus, the pluralism is positively motivated through considerations regarding the purpose that concepts of actual causation fulfil for an agent who uses these concepts. Thus, I aim to provide a functional account of actual causation. The most prominent functional account of causation has been provided by James Woodward (2014). According to Woodward, a functional account "takes as its point of departure the idea that causal information and reasoning are sometimes useful or functional in the sense of serving various goals and purposes that we have" (693). The focus of Woodward's interventionist theory is one particular kind of purpose: manipulation. Woodward argues that a functional account "then proceeds by trying to understand and evaluate various forms of causal cognition in terms of how well they conduce to the achievement of these purposes. Causal cognition is thus seen as a kind of epistemic technology—as a tool—and, like other technologies judged in terms of how well it serves our goals and purposes" (693f).

Woodward's functional approach has both descriptive and evaluative elements. It has descriptive elements because it is based on assumptions about how we actually use causal concepts. It has evaluative elements because it involves an assessment of whether causal concepts facilitate our goals. Moreover, the metaphor of causal cognition as a technology suggests that we can also improve that technology. In that sense a functional account can be understood as pursuing a kind of conceptual engineering (Cappelen, 2018).

One way to employ a functional account is to aim at explaining the distinction between causal and non-causal relations. Such an account has been put forward, for example, by Cartwright (1979) who argues that we need to distinguish causal relations from mere correlations in order to identify effective strategies. For example, it may well be the case that holding a college teachers' life insurance policy correlates with longer lifetimes. However, purchasing a life insurance policy is not an effective strategy for lengthening one's life because it is not a cause of a longer lifetime.

I take my functional account to make a contribution to the more recent project of establishing useful distinctions *among* causes. Many have argued implicitly (e.g. Woodward (2003)) or explicitly (e.g. Hitchcock (2003, 2007b)) that such useful distinctions can be drawn. Woodward, for example, distinguishes a range of causal notions, including direct, total, contributing, and actual causation. *Pace* Woodward, however, I will argue that it is a mistake to list actual causation as one notion among the other notions. Instead, I argue that we need to distinguish between total, path-changing, and contributing actual causation as detailed above.

Another important difference to existing functional accounts is that I do not take our intuitions associated with actual causation to be explained solely by the function it plays in contexts of intervention. In the law the function of this concept is to indicate responsibility for some harm. Most interventionists seem to be aware of this function—yet they spell out the role that the concept plays in our cognitive architecture by referring to interventions. Thus, there seems to be an underlying assumption that intervention and responsibility are in some sense closely related. Hitchcock and Knobe's (2009) account makes this assumption explicit and identifies the ascription of responsibility as one kind of intervention. I will address this assumption more specifically and highlight cases of causal reasoning where intervention and responsibility come apart. I take these cases to indicate potential limitations of a purely interventionist functional account of actual causation.

The second central theme of this thesis (along with pluralism about actual causation) is the context-sensitivity of concepts of actual causation. Context-sensitivity means that if a factor is an actual cause of an event in some context, then there may be other contexts where this is not the case. There are two ways in which context-sensitivity has featured in debates about actual causation. First, actual causation has been related to considerations of normality. The idea is, roughly, that we typically identify only those factors as actual causes that are abnormal. For example, we typically identify a short-circuit as an actual cause of a fire but not the presence of oxygen. This is related to the fact that normally we inhabit places where oxygen is present and short-circuits do not occur. But of course there are contexts where the situation is reversed. If the fire occurs in a scientific experiment that involves a short-circuit in an evacuated chamber, then the presence of oxygen *is* abnormal and, thus, will be identified as actual cause.

Second, most definitions of actual causation that will be discussed here make explicit reference to an underlying causal model. These definitions typically require that the underlying causal model be an appropriate representation of the target system. However, there are no hard and fast criteria of what an apt causal model is. Moreover, what counts as an appropriate causal model in one context may not be acceptable in another context. In particular this concerns whether a causal model represents all the causal structure that is relevant for evaluating a situation and whether it involves scenarios that are not to be taken seriously because they are too far fetched. A causal model that describes the death of the flowers in the municipal gardens as depending on whether the Queen of England waters them may be rejected on these grounds: the scenario that the Queen waters the flowers is just too unlikely to be included in an apt representation of the situation. It should be clear, however, that there may be contexts where the Queen's watering the flower *is* a relevant scenario.

In accordance with the idea of a pluralism about actual causation I will distinguish two kinds of context-sensitivity. Context-sensitivity$_1$ concerns our willingness to entertain counterfactuals involving the occurrence of certain individual events. Our willingness to accept oxygen as an actual cause of fire, for example, depends upon whether we take the absence of oxygen to be a salient scenario. This kind of consideration can also be understood as related to the feasibility of particular interventions: oxygen is usually taken to be a background condition—and legitimately so—because there are many other strategies for preventing fire that are more feasible than preventing the presence of oxygen.

Context-sensitivity$_2$, by contrast, concerns our willingness to consider the occurrence of certain events, *given that* certain other events occur or do not occur. Our identifying the assassin in training as an actual cause of the victim's death, for example, depends on considerations of this kind. He is an actual cause because in a situation where he does not pull the trigger of his gun it is still a serious possibility that the supervisor fails to intervene and that the victim survives as a result. This kind of consideration can also be understood as related to the feasibility of *combinations* of interventions. Do we think there is a possible combination of interventions that would save the victim? Context-sensitivity$_2$ is important, for example, in cases of preemptive prevention. In preemptive prevention an event *c* prevents another event *e* that would have been prevented by another event *d* that would have occurred in the absence of *c*. In these cases our identifying *c* as actual cause depends on how willing we are to take seriously the scenario that *d* fails to occur in the absence of *c*.

The relevance of norms in the debate on actual causation may come as a surprise from the perspective of law, where the notion of actual causation has its origin. Here actual causation is traditionally construed as a notion that is not context-sensitive. I will argue that my disambiguation between context-sensitivity$_1$ and context-sensitivity$_2$ will shed new light on the feasibility of such a norm-free notion of actual causation. More specifically, context-sensitivity$_1$ may to a certain degree be eliminated from the legal inquiry regarding causes. Richard Wright (1985) argues that legal inquiry is concerned with the question whether a particular instance of tortious behaviour stands in a causal relation to the harm. But whether a particular behaviour is tortious (that is, abnormal according to some legal norm) is not part of the causal inquiry. This argument, however, does not extend to context-sensitivity$_2$. Whether, for example, we take the non-use of a defective brake to be an actual cause of an accident depends on whether we are willing to construe the case as an instance of preemptive prevention and this, in turn, depends on whether we take seriously the possibility of certain combinations of events.

Another question concerns how we should account for norm-related considerations in the framework of causal models. One way that has been popular among a wide range of contributors is to invoke a distinction between default and deviant values. The idea is that variables normally (in one or a mixture of its many possible senses) take on their default values whereas the deviant values represent abnormalities. Thomas Blanchard and Jonathan Schaffer (2017) have advanced a powerful argument against defaults. They argue that standard cases that suggest a need for incorporating the default/deviant distinction result from problematic assumptions about the underlying causal models. Rather than extending the formalism of causal models by the default/deviant distinction, we should provide more careful models that do not represent far-fetched scenarios in the first place. I largely agree with Blanchard and Schaffer's argument as a response to standard arguments for defaults. However, I will argue that there are situations where Blanchard and Schaffer's argument has problematic consequences. Norms affect our causal judgements. Thus, when agents disagree about norms they may subscribe to conflicting judgements of actual causation. In such cases it is better to represent the disagreement as a disagreement about default and deviant values rather than as a disagreement about appropriate model decisions. The reason is that if disagreement is confined to defaults, the agents may still agree on an underlying causal model. This is particularly relevant in situations with complex causal structure where disagreement may also arise for epistemic reasons.

The remainder of the Introduction is structured as follows. In section 1.1 I will specify what I take actual causation to be. In section 1.2 I will clarify the relation of actual causation to other causal concepts that have been proposed within the framework of interventionist causal models. In section 1.3 I will relate my own contribution to existing strands in the literature on causation. In section 1.4 I will provide a brief overview of the chapters of this thesis.

1.1. What is Actual Causation?

I will argue that there is a plurality of concepts of actual causation. Nevertheless, we need to say why the concepts fall into the category of actual causation—as opposed to causation *simpliciter*. The examples provided earlier suggest that there are three important features of actual causation.

First, actual causation entails that a certain cause event and a certain effect event occur. When we ask for the actual cause of this particular person's death, we ask for information about the circumstances under which this particular person died. An answer to this question will (at least implicitly) refer to certain general causal claims—such as the fact that, generally, persons with the bodily constitution of the victim die if they ingest a certain amount of poisonous substance. But the point is that the claim of actual causation is not exhausted by such claims: we need to know whether this particular person did in fact ingest a sufficient amount of the poisonous substance.

Second, a claim of actual causation entails that the effect event is in fact brought about by the cause event. There is no consensus on what 'bringing about' means. Nevertheless, it is safe to say that whether one event brought about another event in a particular situation depends on the specifics of that situation. For example, it is a feature of the particular situation that the assassin in training preempts the supervising assassin. In other contexts it may well be the case that the supervisor kills the victim and the trainee is preempted, or both kill the victim at the same time.[4]

Third, actual causation typically involves privileging the cause as particularly salient. We identify the assassin's pulling the trigger of the gun as actual cause of the victim's death but not the absence of a potential gust of wind that could have deflected the bullet. Backgrounding the potential gust of wind is a contextual matter. There may well be instances of assassination where the absence of wind is a salient detail.

Many authors identify actual causation with *singular causation*.[5] Singular cau-

[4] We will see that this kind of information is often even programmed into the causal models. A causal model in which the value of the supervisor variable depends on the value of the trainee value, for example, reflects the the fact that the supervisor waits as a backup in this particular situation.

[5] Often the claim of identity is made implicitly as in Woodward (2003) who examines "the relationship between [type-causal claims] and claims involving actual-, singular-, or token-causal claims" (75), for further examples see Baumgartner and Fenton-Glynn (2013); Baumgartner (2013); Kutach

sation is taken to be a relation between token events as in 'the fire in the Grenfell Tower, London on 14 June 2017 was caused by a malfunctioning fridge' and it is contrasted with causal laws that relate types of events as in 'malfunctioning fridges cause fires.'

It is true that considerations of actual causation most naturally arise with regard to token events. Actual causation depends on which token events occur in a particular situation. For example, the prolapsed disk being an actual cause of a particular person's back pain depends on whether this particular person has a prolapsed disk. However, it is not true that actual causation essentially concerns token events, which means that actual causation is *not* to be equated with singular causation. The reason is that we can make generalized claims about which *kinds of* events occur in which *kinds of* ways. Let me explain. Looking at a population of back pain patients, for example, we can make generalized claims about how many of these patients' back pain has actually been caused by prolapsed disks. Such claims are not token claims. For they concern not a token instance of back pain.[6] But neither are they simply claims of potential causation. For they concern not merely what could be the cause of these patients' back pain but what actually did cause it.

We can also make generalized claims about the way in which events occur. Suppose the assassin in training and his supervisor follow certain general instructions for assassin education, according to which a trainee is to be accompanied by a supervisor who waits as a backup. This particular sequence of token events can then be understood as an instance of a general scheme in which the trainee preempts the supervisor. There are also more mundane cases of generalized preemption.

(2013); Halpern and Hitchcock (2015); Halpern (2016); Blanchard and Schaffer (2017); Fenton-Glynn (2015); Menzies and Beebee (2019).

[6] Is the fact that three patients have back pain because of three prolapsed disks not itself an instance of token causation? From the interventionist perspective that will be central to this the following discussion, such a coarse individuation scheme seems to be difficult to motivate. What matters from the perspective of the doctor is that there are three different patients and, therefore, three medical interventions have to be performed.

Consider the relation between constituents of complex electric circuits. Under circumstances of normal operation, for example, standard electricity supply preempts emergency power supply.

A key advantage of allowing generalized actual causation is that it explains why we are interested in actual causation from the perspective of intervention. From the perspective of an intervening agent strictly singular causal relations are (or should be) of limited interest, especially when they concern events in the past, because we cannot change the past. Generalized claims of actual causation, by contrast, can be extremely useful from the perspective of an intervening agent. Knowing, for example, that assassins in training are typically accompanied by supervising assassins who intervene only if the trainee fails can be quite useful for an intervening agent whose goal it is to save the victim.

Most examples of actual causation that I have discussed so far concern causes and effects in the past—as indicated by 'c caused e', in the *past tense*. These are the most intuitive examples. One reason is that relations of actual causation depend on how events actually occur. And it is often much easier to evaluate this in retrospective. Moreover, the term 'actual' suggests that actual causation describes events that have already occurred or are currently occurring, while events that may occur in the future could still turn out differently.

However, I will not take actual causation to be so confined. That is, actual causation may well describe a relation between events that will occur in the future. Again, a key advantage of allowing actual causation to concern sequences of events in the future is that it explains why we are interested in actual causation from the perspective of intervention. This is because only events that will occur in the future can be under the control of intervening agents.

Earlier I have argued that a natural way to understand the term 'actual' is to contrast it with 'potential.' But this contrast has to be taken with caution if applied to sequences of events in the future. If we take future relations of actual causation to be under control, this entails that the circumstances can be changed such that the events may or may not occur. In this sense future events are potential rather than actual. If this is the case, what does it mean to talk of actual events in the future? Actual events in the future are those events that will occur unless an intervening agent prevents them. In this sense the trainee's pulling the trigger of his gun will be an actual cause of the victim's death, unless an agent will have intervened on the trainee.

Actual causation is more accurately described as a *backward-looking concept*, as suggested by Hitchcock (2017). Backward-looking does not mean that actual causation is necessarily concerned with events in the past. Instead it reflects the idea that actual causation describes "effect-backward reasoning." This is the kind of reasoning we employ if we ask for the causes of a particular kind of effect and is contrasted with "cause-forward reasoning" (Hitchcock, 2017, 118), where we ask for the effects of a particular cause.

1.2. Actual Causation and other Causal Concepts

The purpose of this section is a preliminary clarification of the relation between actual causation and other causal concepts. At the same time, this section shall illustrate why actual causation is a crucial topic for functional accounts of causation. One potential objection to my description of actual causation so far is that the resulting problems are not problems that are to be solved by a theory of causation. Instead they rather seem to concern the pragmatics of causal reasoning in particular applications, such as explanation, intervention and the ascription of responsibility. Another worry is that, if the aim of my account is to draw distinctions *among* causes rather than between causes and non-causes, then it does not seem to be dealing with the philosophical question of what causation fundamentally is.

In a sense the objections are exactly right. The account that I propose here already assumes that actual causation and other concepts of causation can be spelled out with reference to some kind of counterfactual conditional. At the same time, however, I think that from a philosophical point of view there is much more to be said about causation than what would be provided by a minimal theory that captures the difference between causal and non-causal relations. In particular, there is much more to be said from the perspective of a functional account that tries to elucidate why we entertain certain forms of causal reasoning.

Arguably, the most controversial feature of actual causation is the distinction between salient causes and mere background conditions. There are many theorists who do not hesitate to discuss problems associated with preemption and other forms of redundancy. But even among these theorists most are highly critical of the distinction between causes and background conditions (see e.g. Lewis (1973a); Hall (2004), a more detailed discussion will be provided in Chapter 2).

Let me explain why I think that the distinction between cause and mere background condition is so important and needs to be covered by a functional account. According to Woodward, X is a *contributing cause* of Y relative to some variable set $\mathcal{V}$ if and only if (1) there is a path of direct causes leading from X to Y and (2) there is an intervention on X that changes the value of Y if all variables in $\mathcal{V}$ that are not part of this path are held fixed at some value (Woodward (2003), 55, for a more detailed discussion of this definition see Chapter 4). Note that this definition explicitly refers to a particular causal model that is constituted by a set of variables $\mathcal{V}$.[7] Moreover, note that for X to be a contributing cause of Y it is sufficient that an intervention on X leads to a change in Y given that the other variables in $\mathcal{V}$ (that do not lie on the path between X and Y) are held fixed at *some* value. This value does

[7] Chapter 3 will give a detailed exposition of the causal models framework.

not have to be the value that reflects the actual state of the target system or even a state that the target system is likely to be in. As a consequence, decisions about which variables are part of $\mathcal{V}$ and decisions about the variables' possible values have significant influence on whether X is a contributing cause of Y.

Here is an example (taken from Statham (2017)). Usually the consumption of bottled water (BC) is not considered to be a cause of heart disease (HD). In a causal model that consists of these two variables, there would be no edge leading from BC to HD because there is no intervention on BC that changes the value of HD. However, consider the possibility that water reacts with plastic bottles in a way that produces chemical X, which causes heart disease if consumed. Figure 1.1 shows a model of the situation, where XW represents whether the chemical reaction occurs and XC represents whether the dangerous chemical is consumed. According to this causal model, water consumption *is* a contributing cause of heart disease. For if we set $XW = 1$, then there is an intervention on BC that changes the value of HD. Thus, Woodward's definition yields the result that water consumption is not a cause of heart disease according to a model with $\mathcal{V}_1 = \{BC, HD\}$, but that it *is* a cause according to a model with $\mathcal{V}_2 = \{BC, XW, XC, HD\}$.

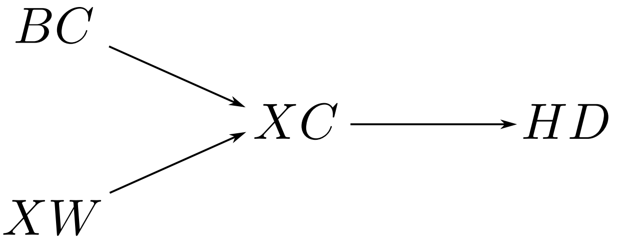

Figure 1.1.: Is consumption of bottled water a contributing cause of heart disease?

Whether XW should be included in $\mathcal{V}$ depends on whether $XW = 0$ is a background condition that can be taken for granted, or whether $XW = 1$ is a scenario that is to be taken seriously. The variable-relativity of Woodward's definition, thus, makes the notion essentially dependent upon context-sensitive considerations regarding background conditions.

In a response to Michael Strevens's review of *Making Things Happen*[8] Woodward also offers a definition of contributing cause that is not relative to a set of variables:

> "X is a contributing cause of Y *simpliciter* (in a sense that isn't relativised to any particular variable set $\mathbf{V}$) as long as it is true that there exists a variable set $\mathbf{V}$ such that X is correctly represented as a contributing cause of Y with respect to $\mathbf{V}$" (Woodward (2008), 209, emphasis original).

This definition can be understood in two ways, depending on what variable sets $\mathcal{V}$ Woodward admits. If Woodward is only referring to the restrictive variable sets (that draw a line between background conditions and factors that are to be taken seriously) then this definition still essentially involves context-sensitive considerations. However, if Woodward means to include all possible $\mathcal{V}$, including cases like the dangerous chemical in the bottled water example, then the notion may not be context-sensitive. However, note that this definition would be extremely permissive. If there is just one causal model according to which a change in C leads to a change in E, then C is considered a contributing cause of E. Woodward is aware of the permissiveness (even of the non-derelativised notion of contributing cause). He argues that

> "the bare claim that X causes Y is not very informative. From the perspective of a manipulability account, what one would really like to know is not just whether there is some manipulation of (or intervention on) X that will change Y; that is, whether it is true that X causes Y. One would also like to have more detailed information about just which interventions on X will change Y (and in what circumstances) and how they will

[8] Michael Strevens criticizes the variable relativity of Woodward's definition for reasons that are different from those discussed here. For a discussion of Strevens's (2007) objection see Woodward's defense (2008), Strevens's reply (2008) and McCain (2015) and Statham (2017) who argue that Strevens's criticism is based on a misunderstanding of Woodward's definition of intervention.

> change Y" (Woodward, 2003, 66).

The more specific information is, of course, provided by the structural equations of the causal model. But Woodward also takes this as a motivation for defining useful distinctions *among* causes, such as those suggested in Woodward (2010).

The crucial point here is that it seems like Woodward's framework offers two options. The first option is that the notion of contributing cause is relativised to a set of variables $\mathcal{V}$ and, thus, essentially depends on the cause/background distinction. This notion seems to be sufficiently restrictive in order to be informative from the perspective of an intervening agent with particular pragmatic goals. The second option is that the notion of contributing cause is derelativised but extremely permissive. In order to provide a theory of causation that is informative from the perspective of an intervening agent one then has to provide further, more specific, contextual information such as provided by claims of actual causation.

Georgie Statham (2017) argues that we should not derelativise. She argues that "[t]o the extent that we don't know about all the causal systems that exist in the universe, it [...] becomes impossible to be sure that any given causal claim is false." She concludes that "it seems preferable to just claim that X is a cause of Y relative to a particular [variable set]" (899). However, I think that it is preferable to assume that there is some such highly permissive and derelativised notion of causation. There are contexts where such a notion of cause is relevant, for example, as a constraint on fundamental physical theories. At the same time, however, I emphasize that a functional theory of causation such as the one that I will provide here needs to address also other notions of causation that are richer in pragmatic content than this minimal notion of causation. Otherwise, we will not be able to explain why agents are interested in causal claims, for example, in contexts that concern manipulability and responsibility.

Looking at concepts of *actual causation* is particularly promising in this regard.

Concepts of actual causation, as I have described them above, involve specific contextual information that is relevant, for example, from the perspective of an intervening agent. This information concerns the actual state of the kind of target system under consideration. Therefore, hypothetical scenarios such as the possibility of dangerous chemicals in the heart disease are largely irrelevant, unless we are dealing with a situation where the rare chemical reaction occurs.

I have argued for the relevance of actual causation by contrasting it with other causal concepts in Woodward's interventionist theory of causation because this theory will be relevant in the following discussion. However, I think that the argument extends to other accounts of causation. Consider, for example, process theories of causation as defended by Wesley Salmon (1984) and Phil Dowe (2000). According to these theories, the defining criterion of a causal process is the transmission of a mark or a conserved quantity. However, Hitchcock (1995) argues that this criterion fails to account for explanatory relevance. For example, "John Jones avoided becoming pregnant during the past year, for he has taken his wife's birth control pills regularly, and every man who regularly takes birth control pills avoids pregnancy" (Salmon, 1971, 34).[9] The problem for process theories is that the criterion of transmission is too permissive in order to exclude cases like these. John Jones's taking birth control pills is a causal process, however, not one that is relevant for his not becoming pregnant. So, causal process theories may capture a certain minimal concept of causation. But they are not sufficient to explain why causation serves the function that it does, for example, in contexts of causal explanation. The basic concept of causal process would have to be complemented with more specific notions in order to give a satisfactory functional account.

[9] Explanatory relevance is one central problem for Hempel and Oppenheim's (1948) deductive-nomological theory of explanation, which states that an explanation deduces the explanans from a set of initial conditions using at least one general law.

1.3. The Relation to other Approaches to Causation

Other Accounts of Causation

In this section I will put my own contribution into relation with existing strands in the literature on causation—both beyond and within the tradition of counterfactual conditionals. There is range of approaches that make different assumptions about what causation fundamentally amounts to. In addition to the counterfactual accounts that will be at the centre of my discussion, there are, first, regularity accounts (Hume (1777/1975); Mill (1843/1882); Mackie (1974); Baumgartner and Falk (forthcoming)). According to these accounts, causation ultimately comes down to underlying laws of nature that need to be distinguished from merely accidental generalizations. Second, there are probabilistic accounts. According to these accounts, causation ultimately amounts to an increase of the probability of the effect (Reichenbach (1956); Suppes (1970); Cartwright (1979); Skyrms (1980); Eells (1991)). A central concern of these theories is to discern causal relations from spurious correlations, such as those resulting from a common cause. Third, there are accounts that define causation in terms of processes, as discussed in the foregoing section.

I will focus on the idea that actual causation can be spelled out in a useful way in terms of counterfactual conditionals. The idea that causation is closely related to such counterfactual conditionals goes back at least to David Hume (1777/1975)[10] and has been central to philosophical debates since the pioneering work of David Lewis (1973a). One straightforward way of implementing counterfactual conditionals would be to require that effects depend counterfactually on their actual causes. However, this criterion seems to fail with regard to redundancy. Redundancy means that in the absence of the cause there are other factors that sustain the effect, which means that the effect does not depend counterfactually on the cause.

[10] Hume takes this to be equivalent to his regularity account, see Chapter 2.

Proponents of counterfactual theories have put forward a range of accounts of the problem of redundancy. Lewis's (1973a) original approach requires that effects be connected to their actual causes by a chain of counterfactual dependence. Later Lewis (1986a) suggested that causation comes down to quasi-dependence, a notion that appeals to the idea of counterfactual dependence as well as the intuition that actual causation is an intrinsic relation. Finally, Lewis (2000; 2004) proposed an account of causation as influence. Here the key idea is that the specific way in which the effect occurs depends counterfactually on the way the cause occurs. All these accounts have been confronted with counterarguments, which will be discuss in Chapter 2.

Even though my focus will be on accounts according to which actual causation is usefully spelled out in terms of counterfactual dependence *under certain contingencies*, I take the resulting pluralism about actual causation to be of a more general nature. That is, I do think that the pluralism is most conveniently defended within the framework of causal models. However, I also think that the pluralism affects other theories of actual causation as well. Lewis's original ancestral dependence account, for example, involves a distinction between causation and causal dependence that is analogous to my distinction between total and path-changing actual causation. Likewise, Lewis's quasi-dependence account seems to suggest that we can distinguish between dependence and quasi-dependence and, presumably, also within the framework of the influence account one can distinguish different kinds of influence.[11]

Other Projects in the Philosophy of Causation

My pluralist theory about actual causation is motivated by a functional approach. Following Woodward (2014; 2015), such a functional approach can be contrasted

[11] It is an interesting question whether these distinctions extend to other theories that do not use counterfactual conditionals. But this question is beyond the scope of my discussion.

with contributions to the *metaphysical project* of causation which has the aim to determine the "truth-makers" or "grounds" of causation.[12] Further questions that are relevant in this project concern whether causation is among the fundamental constituents of reality, and what the true relata of the causal relationship are. Moreover, contributors to this project tend to emphasize that the analysis should not be spoiled by merely pragmatic considerations. I will not have much to say about the metaphysical project. For the most part I will take it for granted that causation, and actual causation more specifically, can be spelled out usefully in terms of counterfactual conditionals in one way or another. Thereby, I do not, however, mean to commit to the metaphysical claim that causation reduces to counterfactuals. Moreover, by saying that causation can be *usefully* spelled out in terms of certain kinds of counterfactuals I explicitly endorse the relevance of pragmatic factors.

According to Woodward (2014), two further projects within the philosophy of causation are the *descriptive project* and the *fit with physics project*. Proponents of the descriptive project, according to Woodward, "attach considerable importance to constructing accounts whose aim is to describe or reproduce (what they take to be) the causal judgments of "ordinary folk" regarding various scenarios" (692). Woodward refers to mostly intuition-guided contributions such as those discussed in Collins et al. (2004). But presumably studies on causal reasoning performed by experimental philosophers or empirical psychologists (Walsh and Sloman (2005); Knobe and Fraser (2008); Lombrozo (2010)) are also part of this project. A purely descriptive account is surely to be distinguished from the functional project. For whether a certain concept fulfils a particular function is an evaluative question. However, I will not take the distinction to be so clear cut. The reason is that I take my functional account to be informed by descriptive claims. This seems to be in agreement with contributors such as Hitchcock and Knobe (2009) and Sytsma et

[12] Woodward identifies Tooley (1977), Armstrong (1983), and Bird (2005) as instantiations.

al. (2012) who take their functional accounts to be informed by empirical studies on causal judgement.

The fit with physics project, according to Woodward, is the somewhat vaguely defined project focussing "on issues having to do with the relationship between causal claims [...] and what is imagined by some philosophers to be "fundamental physics"" (2014, 692). Woodward identifies sceptical arguments along the lines of Hartry Field (2003) and Barry Loewer (2009) as examples for this project.[13] Here, again, the contrast with the functional project does not seem to be as clear cut. Field's and Loewer's arguments seem to draw at least part of their motivation from a perceived lack of usefulness of causal notions in physics. And Mathias Frisch argues against the scepticism of Field and Loewer and provides a "functional defense of causal reasoning [...] in physics" (2014, 11).[14] I will not address the fit with physics project. Nevertheless, I think that my functional analysis of actual causation might as well be usefully applied to causal reasoning—at least—in experimental physics. For example, do physicists try to identify the actual causes of measurement results and phenomena or is it sufficient to find potential causes? Sometimes measurement deviations or surprising phenomena motivate a systematic search for underlying actual causes, such as in the discovery of the Cosmic Microwave Background Radiation. Sometimes, however, it is sufficient to point to potential causes, for example, when the deviation is within the limits of the errorbars. Moreover, selective causal reasoning seems to be particularly relevant in experiments. In fact, it has been argued that experiments are *designed* to enable selective causal reasoning. According to Norwood Russell Hanson, it is the very point of skilful experimentation "to bring together a cluster of theoretical considerations in a single, tersely-expressed hypothesis [and] to torture it in an experiment, each phase of which keeps everything

[13] The contributions in Price and Corry (2007) are also concerned with this project.

[14] See Frisch (forthcoming) for an argument that the functional and the fit-with-physics project are not necessarily distinct projects.

constant except one set of factors [...]" (1958, 67).

1.4. Overview of the Chapters

The thesis has three parts. Part I will provide the background for my pluralist account. In Chapter 2 I will introduce two central problems for theories of actual causation: the problem of redundancy and the problem of selection. In Chapter 3 I will review accounts that address these problems by employing the framework of causal models. In Part II I will develop my pluralist account. In Chapter 4 I will argue from an interventionist perspective that we need to distinguish a range of different concepts of actual causation. In Chapter 5, I will argue that we also need a pluralist account with regard to the function of these concepts. In Part III I will explore consequences of the pluralist account with particular regard to the context-sensitivity of actual causation. In Chapter 6 I will show that attempts to provide a principled approach to actual causation in the law face difficulties. In Chapter 7 I will provide a new argument for incorporating a distinction between default and deviant values into the formal framework of causal models. In Chapter 8 I will summarize the main results and provide an outlook.

Part I – A Unified Account of Actual Causation?

Chapter 2 – The Problems of Redundancy and Selection

In Chapter 2 I will introduce two challenges for theories of actual causation: the problem of redundancy and the problem of selection. There are four kinds of redundancy: symmetrical overdetermination, early preemption, late preemption, and trumping. I will also review four kinds of approaches for solving the problem of redundancy that retain the basic idea of counterfactual dependence. First, I will discuss the idea that effect events are to be construed in a very fine-grained way. Second, I will review Lewis's account of transitive causation. Third, I will address attempts to exploit the intrinsic nature of causation. Finally, I will address Lewis's idea that causation is to be spelled out in terms of influence. I will point out where these accounts encounter difficulties in order to prepare the discussion of the following chapters.

I will then turn to the problem of selection. I will point out what exactly the problem of selection is by identifying what I call *Mill's challenge*: selection is capricious and not justified. Mill's challenge can be met by showing that there are principles that guide causal selection and that these principles are justified. I will address two kinds of approaches and assess whether they meet Mill's challenge. First, there are contextual-variable accounts. According to these accounts, the problem of selection arises if some contextual variable (e.g. causal field, contrast class, framework) is left implicit. Second, I will discuss normality-based accounts. According to these accounts, we tend to select those factors that are abnormal, where normality can be understood in a range of different ways, including descriptive as well as prescriptive senses. I will argue that normality-based accounts are contextual as well, yet they are more informative than the contextual-variable accounts because they can explain why causal selection is a largely stable phenomenon and why we legitimately select some causes rather than others.

Chapter 3 – Causal Models for a Unified Account?

In Chapter 3 I will review existing attempts to provide a unified account of actual causation in the framework of causal models. After a brief introduction to the formalism of causal models I will first address a series of accounts that define actual causation as dependence given that certain variables are held fixed at their *actual* values. These accounts provide a successful treatment of cases involving early preemption but face problems in cases involving symmetrical overdetermination.

I will then address accounts that define actual causation as dependence given that certain variables are set to *non-actual* values. These accounts succeed with respect to early preemption and symmetrical overdetermination. However, they do so at the price of providing a notion of actual causation that seems to be too permissive in order to capture certain causal intuitions. In particular, they face the Problem of Isomorphism. The Problem of Isomorphism describes instances where two example cases seem to have isomorphic causal models but give rise to conflicting causal judgements. This is a problem for accounts of actual causation that are based solely on the structural features of causal models. Therefore, a number of authors have suggested extending the formalism of causal models by a distinction between default and deviant values. I will introduce such accounts and will illustrate how the distinction is supposed to solve the Problem of Isomorphism.

Part II – Pluralism about Actual Causation

Chapter 4 – Pluralism about Actual Causation

In Chapter 4 I will advance a pluralist account about actual causation. I will argue that we need to distinguish total actual causes, path-changing actual causes, and contributing actual causes. Total actual causation involves a straightforward counterfactual dependence of the effect on the cause. The notion of path-changing actual causation involves counterfactual dependence given that certain values are fixed at their actual values. Finally, contributing actual causation involves counterfactual dependence given that certain variables are set to non-actual values. The pluralist account is supported by two lines of argument. First, I will provide a set of toy examples that raise difficulties for unified accounts and I will show that my pluralist theory accounts for them. Second, I will provide a functional justification. An important function of reasoning in terms of actual causation is to indicate suitable targets of intervention. That is, claims about the actual causes of an undesired effect inform agents about suitable interventions that help avoid this effect. But if this is the case, then these claims should *at least* tell the agent whether one or several interventions have to be applied and whether the additional interventions need to target other aspects of the situation that are currently actualized. Moreover, I shall distinguish two kinds of context-sensitivity. Context-sensitivity$_1$ concerns the normality or typicality of an individual variable's possible values. Context-sensitivity$_2$ concerns considerations regarding possible violations of the model's structural equations.

Chapter 5 – Responsibility and the Limits of Interventionism

In Chapter 5 I will examine the assumption that the function of concepts of actual causation is to facilitate intervention. More specifically, I will identify two kinds of situations where interventionist explanations of the function of actual causation face difficulties. First, interventionists have difficulties to explain our interest in certain selective claims of total actual causation. Interventionists argue that we identify norm-violating factors as actual causes because they are particularly suited as targets for intervention. However, there are cases where token claims of total actual causation clearly do not correspond to such suitable targets for intervention. Second, in cases of redundancy the difference between symmetric overdetermination and late preemption is (at least sometimes) difficult to capture from an interventionist perspective. I will argue that where the interventionist perspective reaches its limits, incorporating responsibility helps us make sense of judgements of actual causation.

Part III – Consequences

Chapter 6 – Actual Causation in the Law

In Chapter 6 I will discuss the concept of actual causation in the law. In the law the notion of actual causation goes back to the American Legal Realists of the early 20th century. These were concerned with delineating the factual elements of a legal inquiry from those elements that depend upon norms and policy. "Actual causation" is meant to refer to the factual elements that can be approached in a principled way, while the norm-related and context-sensitive elements are described by the notion of "proximate cause." This, however, seems to stand in conflict with the more recent use of the term "actual causation" in the causal models literature, where many contributors take it to be a context-sensitive notion. In this chapter I will disentangle these apparently conflicting takes on the context-sensitivity of actual causation. There seem to be two possible solutions. The first possible solution is that the causal models literature (or large parts of it) has been right in arguing that actual causation is (at least sometimes) context-sensitive. Then the literature on causal models should provide us with arguments that undermine the Legal Realist's project. Alternatively, it could be the case that a principled account *is* possible and that contributors to the causal models literature simply have been misguided in using the term actual causation in order to refer to context-sensitive aspects of causal reasoning. I will argue that either conclusion would be too quick. Context-sensitivity$_1$ can indeed be excluded from the lawyer's causal inquiry. However, it is unlikely that the Legal Realist is able to exclude context-sensitivity$_2$. I thus use the distinction introduced in Chapter 4 in order to shed new light on the notion of actual cause in legal inquiry.

Chapter 7 – Causation and the Problem of Disagreement

In Chapter 7 I will provide a new argument for incorporating the default/deviant distinction into causal models. There are two ways to formalize context-sensitivity in causal models. Either (i) one can adjust the model such that it reproduces the plausible causal claims directly or (ii) one can enrich the causal structure by introducing a distinction between default and deviant values of variables. Thomas Blanchard and Jonathan Schaffer have argued for strategy (i) putting forward a series of forceful arguments against the distinction between defaults and deviants. I will argue—*pace* Blanchard and Schaffer—that defaults have an important role to play. My argument is based on cases where causal reasoners agree about the causal structure but disagree about the appropriate causal judgements. In this kind of context causal models should be seen as a means of representation that facilitates a clarification of different kinds of disagreement: epistemic disagreement about causal structure and normative disagreement about scenarios that are to be taken seriously.

Part I.

A Unified Account of Actual Causation?

2. The Problems of Redundancy and Selection

2.1. Two Problems for Counterfactual Theories of Actual Causation

In this chapter I will introduce and discuss two problems for counterfactual theories of actual causation: the problem of redundancy and the problem of selection. The discussion of these two problems shall motivate and prepare the accounts of actual causation that will be discussed in Chapter 3. At the same time, this chapter shall support a claim that I have put forward in the Introduction: actual causation is a contextual notion.

The chapter is structured as follows. In section 2.2 I will briefly explain what a counterfactual theory of actual causation is. In section 2.3 I will introduce four kinds of redundancy and explain why they are a problem for counterfactual definitions of actual causation. In section 2.4 I will address four approaches to redundancy that have been suggested and discussed by proponents of counterfactual accounts: fragility, ancestral dependence, quasi-dependence, and influence. I will explain why these accounts have been considered unsatisfactory. I will then turn to the problem of selection. In section 2.5 I will specify what the problem of causal selection is. More specifically, I will identify what will be called *Mill's challenge*: the claim that causal selection is capricious, that is, that there are no guiding principles for selection and that we are not justified to select. In section 2.6 I will discuss contextual-variable accounts of causal selection. According to these accounts, the problem of selection is solved by identifying a certain contextual variable, such as relevant contrast classes. Contextual-variable accounts succeed in explaining the apparent capriciousness of causal selection. However, they do not give guiding principles for fixing the relevant contextual variable. In section 2.7 I will address recent and promising attempts to spell out selection in terms of a concept of normality. I will argue that these accounts are clearly contextual as well—because the notion of normality is contextual. Nevertheless, the notion of normality also shows that there are at least some guiding principles for causal selection and that these principles are sometimes justified.

2.2. Counterfactuals and Actual Causation

A counterfactual conditional is a conditional of the form 'if the bottle had been hit, then the bottle would have shattered.' Both the term 'counterfactual' and this example seem to imply that the antecedent of a counterfactual conditional is false. But in the following I shall presume a broader understanding of counterfactuals that is common among philosophers. According to this understanding, counterfactuals can have antecedents that may be true, as in 'if the bottle were hit, then it would shatter.'[1]

The first explicit counterfactual definition of causation was provided by David Hume in his *Enquiry Concerning Human Understanding* (1777/1975). Unfortunately, however, Hume took the definition to be equivalent to his definition in terms of regularities.[2] For a long time, regularity analyses of causation dominated the philo-

[1] This is different in the psychological literature, where counterfactuals are typically considered to have false antecedents, see Hoerl et al. (2011).

[2] "we may define a cause to be *an object followed by another, and where all the objects, similar to the first,*

sophical debate. Following a systematic analysis of the semantics of counterfactual conditionals in terms of possible worlds (Stalnaker (1968); Lewis (1973b)), David Lewis's seminal article *Causation* (1973a) put counterfactuals on the agenda of the philosophy of causation.

Lewis's original counterfactual account of causation is developed in three stages. We shall look briefly at the first two stages and address the third stage in section 2.4.2. At the first stage Lewis defines a notion of counterfactual dependence as follows: "Let there be two families A and C of propositions $A_1, A_2, ...$ and $C_1, C_2, ...$ If there are true counterfactuals that relate all propositions in family A to propositions in family C such that $A_1 \mathrel{\Box\!\!\rightarrow} C_1, A_2 \mathrel{\Box\!\!\rightarrow} C_2, ...$, then the C's *depend counterfactually* on the A's" (1973a, 561, emphasis original). Here $A \mathrel{\Box\!\!\rightarrow} C$ is the counterfactual 'if it were the case that A, then it would be the case that C.'

Not all instances of counterfactual dependence are causal. Consider the mathematical function $f(x) = ax$. There is a relation of counterfactual dependence between the value of the coefficient a and the slope of the function's graph. But the dependence is clearly not causal. Lewis rules out such instances by requiring counterfactual dependence among propositions that describe the occurrence of events: "The family $e_1, e_2, ...$ of events *depends causally* on the family $c_1, c_2, ...$ iff the family $O(e_1), O(e_2), ...$ of propositions depends counterfactually on the family $O(c_1), O(c_2), ...$" (562, emphasis originial). This is the second stage of Lewis's account.

For example, the bottle's shattering depends causally on its being hit. More precisely, this means that whether the bottle shatters ($O(s)$) or not ($\neg O(s)$) depends on whether it is hit ($O(h)$) or not ($\neg O(h)$). This is captured by two counterfactual conditionals $O(h) \mathrel{\Box\!\!\rightarrow} O(s)$ and $\neg O(h) \mathrel{\Box\!\!\rightarrow} \neg O(s)$. Lewis distinguishes two cases. First, if the bottle is not hit and does not shatter, then the second counterfactual is

are followed by objects similar to the second. Or, in other words, *where, if the first object had not been, the second never had existed*" (Hume (1777/1975, Section VII, emphasis original).

automatically true. In order to establish causal dependence we have to check the first counterfactual. Second, if the bottle is hit and shatters, then the first counterfactual is automatically true and we have to check whether the second counterfactual is true. The second case concerns the actual causal claim that the bottle's shattering was caused by its being hit.

In the following we will look at two particular problems that arise for counterfactual accounts. The first problem is the problem of redundancy. In Lewis's terminology the problem is that in cases of redundancy effects do not depend causally on their purported causes. The problem of selection is that an effect stands in a relation of causal dependence to a much broader range of conditions than those that would typically be called causes.

There are other problems for counterfactual accounts. Lewis's possible-world semantics is based on a similarity relation between worlds. Instead of a principled account, Lewis provides a set of rules that he derives from "what we know about the truth and falsity of counterfactuals" (1979, 43). It is a particular challenge to provide similarity relations that disallow backtracking counterfactuals, such as 'if the bottle had been shattered, then it would have to have been hit.' However, these problems shall not concern us here because they concern counterfactual theories of causation generally and we shall focus on actual causation.

2.3. Redundancy

Redundancy means that in the absence of a cause there are other factors that sustain the effect. This means that the effect does not depend counterfactually on the cause. Thus, redundancy undermines counterfactual dependence as a necessary criterion for actual causation.

Redundancy will be treated here as a threat to counterfactual dependence being a necessary condition for *actual* causation. Does redundancy affect potential causation as well? Suppose counterfactual dependence is taken as a necessary condition for potential causation as follows: if c is a potential cause of e, then there are possible circumstances in which e depends counterfactually on c. This is not undermined by the kinds of redundancy that will be discussed here. Suppose, for example, Suzy and Billy throw stones at a bottle and Suzy preempts Billy in destroying the bottle. One kind of possible circumstance is that Billy does not throw. Then the bottle's shattering would depend on Suzy and, thus, Suzy is a potential cause. Likewise, Billy is a potential cause because one possible kind of circumstance is where Suzy does not throw and, thus, the bottle's shattering depends on Billy's throwing.

This being said, there may be other kinds of redundancy that affect potential causation. Consider, for example, mental causation. Presume some kind of non-reductive physicalism, that is, there are mental properties that are not physical. At the same time, we shall assume that mental properties are strongly dependent upon physical properties, for example, because they are constituted by them. Are mental properties potential causes of other mental properties? According to causal exclusion arguments (see e.g. Malcolm (1968); Kim (1989)), they are not. For if mental properties M are realized by physical properties P and these physical properties P are caused by other physical properties P', then it seems that there is no causal work left for the mental properties M' that correspond to P'. In this sense mental and physical processes can be understood as redundant causes.

In the following discussion we will focus on four kinds of redundancy that affect actual causation: symmetric overdetermination and early preemption, late preemption, and trumping. In symmetric overdetermination two (or more) events c_1 and c_2 cause some effect e and both c_1 and c_2 are sufficient for the effect e. This is a problem for counterfactual definitions because this means that e does not depend counterfactually on either c_1 or c_2. For example, suppose a forest fire was caused by two lightning bolts, l_1 and l_2, such that each of the lightning bolts would have been sufficient for the fire. Then the fire does not depend counterfactually on either of the lightning bolts. Yet we think that both of the two lightning bolts are causes of the fire.

In cases of preemption there are also two (or more) events c_1 and c_2 that would be sufficient for the occurrence of e. But here only one of the events (c_1) causes e while the other event (c_2) is preempted. Preemption comes in two varieties: early and late preemption. An example for early preemption is the following situation, called "Backup":

> "an assassin-in-training is on his first mission. Trainee is an excellent shot: if he shoots his gun, the bullet will fell Victim. Supervisor is also present, in case Trainee has a last minute loss of nerve (a common affliction among student assassins) and fails to pull the trigger. If Trainee does not shoot, Supervisor will shoot Victim herself. In fact, Trainee performs admirably, firing his gun and killing Victim" (Hitchcock, 2001, 276).

Trainee's shot is the cause of Victim's death. But Victim's death does not depend counterfactually on Trainee's shot. For if Trainee had not shot, Supervisor would have shot and Victim would have died anyway. This kind of preemption is called early preemption because the alternative causal process (Supervisor's shooting Victim) is doomed at an early stage, that is, before Supervisor has pulled the trigger of her gun.

The following case is an example of late preemption:

> "Suzy and Billy, expert rock-throwers, are engaged in a competition to see who can shatter a target bottle first. Both pick up rocks and throw them at the bottle, but Suzy throws hers a split second before Billy.

> Consequently Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's had not occurred [...]" Hall (2004, 235).

As in early preemption there are two processes that could cause the bottle's shattering: Suzy's throwing her stone and Billy's throwing his stone. Since only Suzy hits the bottle, we identify her throwing as the actual cause of the bottle's shattering. Yet there is no relation of counterfactual dependence between the bottle's shattering and Suzy's throwing her stone. For Billy's stone would have destroyed the bottle if Suzy had not thrown. The crucial difference to "Backup" is that the only reason why Billy's stone does not shatter the bottle is that by the time the stone arrives at the bottle's original position, the bottle has been destroyed by Suzy's stone. Generally, in late preemption the primary process interrupts the alternative process only through causing the effect. This is different from early preemption where the primary process interrupts the alternative process in some other way and earlier. For example, Supervisor's attack on the Victim is interrupted already at the stage where Supervisor sees that Trainee pulls the trigger.

Finally, here is an example of trumping:

> "Imagine that it is a law of magic that the first spell cast on every day match the enchantment at midnight. Suppose at noon Merlin casts a spell (the first that day) to turn the prince into a frog, that at 6:00 P.M. Morgana casts a spell (the only that day) to turn the prince into a frog, and that at midnight the prince becomes a frog" (Schaffer, 2000a, 165).

There are two processes that could bring about the effect: Merlin's spell and Morgana's spell. According to the law of magic, only Merlin's spell is an actual cause of the transfiguration, because it was cast first. The difference between trumping cases and cases of early and late preemption is that in trumping the alternative process is not interrupted at all and comes to an end just as the preempting process. There are also other, more mundane, instances of trumping. For example, suppose "the major and the sergeant stand before the corporal, both shout "Charge!" at the same time, and the corporal decides to charge" (Schaffer, 2000a, 175). Again none of the two causal processes is interrupted, for the corporal hears both orders. Yet many think that only the major's order is an actual cause of the sergeant's charging because orders of higher-ranked officers trump orders of lower-ranked officers.

Schaffer describes these two cases as instances of trumping *preemption*. But are trumping cases really cases of preemption?[3] In a sense this seems to be wrong because both causal processes run to completion, just as in cases of symmetrical overdetermination. Describing the case as an instance of symmetrical overdetermination, however, seems to neglect the asymmetry between the trumping cause and the trumped cause. In Chapter 4 we will see that both the fact that both causal processes run to completion and the preceived asymmetry are important features of this kind of case. I will argue that the difference between the two factors is not one between a cause and a non-cause but rather a difference between different kinds of actual causes.

2.4. Four Approaches to Redundancy and Their Problems

In this section I will discuss four (kinds of) counterfactual accounts that address cases involving redundancy. But before we turn to these accounts let me briefly note that other accounts also face difficulties with redundancy. First, consider regularity theories. According to Mackie (1974), for example, a cause has to fulfil the INUS criterion, that is, it has to be an **i**nsufficient but **n**ecessary element of an **u**nnecessary but **s**ufficient set of antecedent conditions. This criterion, however, faces difficulties

[3] This has been contested, for example, by Halpern and Pearl (2005) and Hitchcock (2011).

with regard to preemption cases. For both the preempting and the preempted factors fulfil the INUS criterion. A detailed treatment of preemption along these lines is provided, for example, by Richard Wright (1985). I will discuss Wright's account in Chapter 6.

Preemption also raises problems for probabilistic accounts of actual causation (Menzies (1989, 1996); Fenton-Glynn (2015)). According to probabilistic accounts of causation, causes raise the probability of their effects. Yet, consider the following probabilistic version of early preemption. Suppose that this time the assassin in training is not as experienced. In fact, it is the trainee's first mission and it is highly unlikely that he will hit the victim upon pulling the trigger of the gun. His supervisor, however, is reliable. If she decides to intervene, it is highly unlikely that she will miss her target. Yet, the supervisor will shoot only if she sees that the assassin in training fails to pull the trigger of his gun. The trainee pulls the trigger of his gun and, against all odds, kills the victim. As in deterministic early preemption the victim's death is caused by the trainee's pulling the trigger of his gun. And this is so even though the trainee's pulling the trigger of his gun (at least on one reading of the case) *decreased* the probability of the victim's death.

Finally, preemption raises problems for process theories of causation because it leads to so-called misconnection (see e.g. Ehring (2003) and the discussion in Dowe (2004)). Suppose we take transference of a conserved quantity as a sufficient criterion for causation. In the late preemption case this means that Suzy will be identified as an actual cause of the shattered bottle because the momentum of her stone is transferred to the bottle. Yet due to the minute gravitational influence on the bottle there is also a causal process linking Billy's throwing his stone and the bottle's shattering, even though Billy's throwing his stone is clearly not an actual cause. That is, process accounts misconnect Billy's throwing his stone and the bottle's shattering.

Thus, redundancy poses a challenge to the most common theories of causation. I have indicated here that contributors to each of these theories have given accounts of redundancy. In the following, however, we shall focus on the tradition of counterfactual accounts. The reason is that the problems that this tradition faces will be most instructive for the discussion in the subsequent chapters.

2.4.1. Fragility

Consider the Suzy-Billy case. The case describes two processes that compete for being the actual cause of one and the same effect event: the bottle's shattering. But do they really? Billy throws his stone a little later than Suzy and, presumably, from a different angle. Thus, if he had destroyed the bottle, then the bottle's shattering would clearly have been different from the shattering that results from Suzy's stone. Thus, it seems like an undue simplification to say that Suzy's and Billy's throws compete with each other. If we employ a more precise description of the possible outcomes, then there is a particular bottle shattering that depends counterfactually upon Suzy and another particular bottle shattering that depends upon Billy.[4]

The general idea is that, given we find a sufficiently detailed event description, counterfactual dependence as a criterion for causation can be restored. Such details are captured if we employ a fine-grained individuation scheme that specifies an event's exact time and manner of occurrence. These details are then considered to be essential features of the event. Consequently, such an event tends to be fragile in the sense that it could not have occurred at a different time or in a different manner. By contrast, a coarse-grained individuation scheme does not refer to such details. Consequently, events tend to be robust under such an individuation scheme.

However, fragility as a solution to redundancy has two problems. First, it depends

[4] Similar solutions apply to symmetrical overdetermination and early preemption: the fire caused by two lightning bolts is not exactly the same as a fire caused by just one of the lightning bolts, Supervisor's shot will kill the Victim in a slightly different way than Trainee's shot.

upon the contingent fact that the causal processes lead to different outcomes (even if the difference is ever so slight). This need not be so. For example, in cases of trumping the exact manner in which the corporal charges does not depend on the ordering officer's rank (at least it does not need to depend on this).

Second, fragility leads to a profusion of causation, that is, it introduces causal relations where we typically think there are none. Consider the following situation:

> "Boddie eats a big dinner, and then the poisoned chocolates. Poison taken on a full stomach passes more slowly into the blood, which slightly affects the time and manner of the death. If the death is extremely fragile, then one of its causes is the eating of the dinner" (Lewis, 1986a, 198).

This is implausible. A similar worry arises in the Suzy-Billy case. If we apply a sufficiently fine-grained individuation scheme to the bottle shattering, the shattering will depend upon the minute gravitational force exerted by Billy's stone. Consequently, Billy's throwing would count as an actual cause of the bottle's shattering after all. Thus, the problem is that with fragile events we generate causal relations where there are clearly none.

Note, however, the context-dependence of individuation schemes. Suppose the murderer intended Boddie to die quickly and without pain. The murderer may regret that Boddie's dinner caused the death to be so long and painful instead. The problem then is that there are no principled criteria that tell us when to take events to be robust and when fragile. Lewis thinks that finding such principles "may not be a hopeless project, but for the present it is not so much unfinished as unbegun" (Lewis, 1986a, 199).

2.4.2. Ancestral Dependence

Lewis's (1973a) original theory attempts to account for redundancy by defining causation as the ancestral of causal dependence. We shall begin with early preemption.

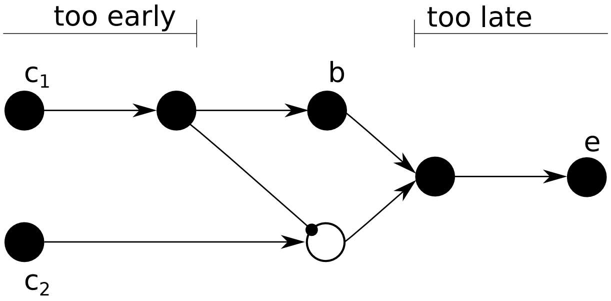

Figure 2.1.: Neuron diagram of Backup.

Figure 2.1 displays a neuron diagram that represents Backup. Neuron diagrams are a graphical tool for representing causal structure like directed acyclic graphs. The key difference is that neurons represent token events that do (full circle) or do not (empty circle) occur. Moreover, the edges carry more information about the relation between the adjacent neurons. Arrows between neurons represent excitatory influence. That is, a neuron is activated if one of its incoming arrows is attached to an activated neuron. An exception are neurons that are (also) under inhibitory influence, represented by edges with a circle instead of the arrow-head. If a neuron is under inhibitory influence, then it is inactive, even if it receives also excitatory input.

In virtue of what difference is Trainee's shot a cause but not Supervisor's shot? According to Lewis (1973a), the key difference is that there is a chain of causally dependent events linking Trainee's actions (c_1) to Victim's death (e), but there is no such chain for Supervisor's actions (c_2). The chain of causal dependence is revealed if we consider an event, b, that lies on the path leading from Trainee's actions to Victim's death: the bullet's propagating from Trainee's gun to Victim. Victim's death depends causally on b and b depends causally on Trainee's actions. Thus, c_1, b, and e constitute a chain of causally dependent events. By contrast, there is no such chain between Supervisor's actions (c_2) and Victim's death (e). Any intermediate event that depends causally on Supervisor's actions (that is, any event that occurs before Trainee's pulling the trigger) does not stand in a relation of causal dependence with Victim's death. Therefore, Lewis refers to the existence of a chain of causal dependence—i.e. a causal chain—as a defining criterion of causation: "one event is a *cause* of another iff there exists a causal chain leading from the first to the second" (Lewis, 1973a, 563).

Note that in order to act as an intermediate event in a chain of causal dependence b has to be chosen appropriately. First, it should not occur too early because the alternative (preempted) process has to be doomed already. Otherwise there is no causal dependence of e on b. In Backup the bullet's being on its way to Victim fulfills this condition because at this stage it is clear that Trainee did not have a last minute loss of nerve and Supervisor did not have to step in. If, in this situation, the bullet would miraculously disappear, there would no longer be a backup process to guarantee Victim's death. By contrast, Trainee's directing his gun at Victim would be an event that is too early. Victim's death does not depend causally on this event, because even if Trainee had not aimed, then Victim would still have died through Supervisor's shot. Second, event b should not occur too late such that it also lies on the path linking c_2 with e. Otherwise b fails to depend causally on c_1. Trainee's bullet's propagating towards Victim fulfills this condition because it depends on Trainee's pulling the trigger. By contrast, victim's heart failure would be an event that is too late—presuming that both Trainee and Supervisor kill by aiming at their victims' hearts.

We shall now turn to late preemption. The problem with late preemption is that there is no event b that would establish a chain of causal dependence between the actual cause and the effect.[5] Consider the Suzy-Billy scenario. There is no event that could establish a chain of causal dependence between Suzy's throwing and the bottle's shattering (Lewis (1986a)). We would need to identify an event that occurs

[5] Symmetrical overdetermination and trumping pose problems for the same reasons.

no earlier than the bottle's shattering. For only due to the bottle's shattering it is the case that Billy's attempted shattering is doomed. But the bottle's shattering is already too late because it lies on both the path originating in c_1 and the path originating in c_2 and, therefore, it does not depend causally on Suzy's throwing. Thus, Lewis's causal chain approach fails to account for late preemption.

Another problem with Lewis's causal chain account arises from the assumption of transitivity. The problem is that there are cases where causation is not transitive. One such case is the "Dog Bite" case (McDermott, 1995). A right-handed terrorist plans to detonate a bomb. However, a dog bites off the terrorist's right forefinger. Therefore, the terrorist uses his left forefinger instead to detonate the bomb. The terrorist's detonating the bomb with the left forefinger depends on the dog bite and the bomb's exploding depends on the terrorist's detonating the bomb with the left forefinger. Yet, we do not consider the dog bite a cause of the explosion. The problem is a mismatch in the chain of dependence: the dog bite influences which finger the terrorist uses to detonate the bomb, but the bomb explodes no matter which finger is used to detonate it.

Another counterexample is "Boulder":

> "a boulder is dislodged, and begins rolling ominously toward Hiker. Before it reaches him, Hiker sees the boulder and ducks. The boulder sails harmlessly over his head with nary a centimeter to spare. Hiker survives his ordeal" (Hitchcock, 2001, 276).

Again there is a chain of causal dependence: Hiker would not have ducked if the boulder had not approached him and Hiker would not have survived if he had not ducked. But clearly the boulder's falling is not a cause of Hiker's survival. The falling boulder is the only reason that Hiker's life was threatened in the first place.

These problems are particularly severe since the assumption of transitivity is also built into the following two accounts. However, for now we shall sideline issues of transitivity. I will get back to them in Chapter 3.

2.4.3. Quasi-dependence

Lewis (1986a) develops the quasi-dependence account as an extension of the original counterfactual account in order to deal with late preemption. The central idea is to appeal to the intuition that causation is an intrinsic relation. Imagine a new situation in which Suzy throws her stone at a bottle and Billy is absent. Suzy's destroying the bottle is an intrinsic duplicate of Suzy's destroying the bottle in the original scenario. Moreover, in the new situation the bottle's shattering *does* depend on Suzy's throwing the stone. If Suzy causes the shattering in this new situation and causation is an intrinsic matter, then it should follow that Suzy causes the shattering in the original situation.

Thus, according to Lewis's updated account, e quasi-depends on c if (i) "in its intrinsic character it is just like processes in other regions (of the same world, or other worlds with the same laws) situated in various surroundings" (1986a, 205) and (ii) in these other processes the counterpart of e depends counterfactually on the counterpart of c. In analogy to the ancestral dependence account, a causal chain is then defined as "a sequence of two or more events, with either dependence or quasi-dependence at each step" (Lewis, 1986a, 205). An event c is an actual cause of event e if there is such a causal chain.

Lewis's definition of quasi-dependence raises a number of questions: what are criteria of similarity between processes? Surely, we will not find exact duplicates of the process initiated by Suzy in circumstances where Billy is absent because if Billy is absent, then also his minute gravitational influences on the process are absent. What exactly are the other "various surroundings"? Surely, we would need a criterion for picking the right surroundings that does not entail a circular reference to intrinsicness. These problems may be solved by a more careful definition (as, for example, the one provided by (Hall, 2004)) and shall not concern us here.

A more serious problem is the following. The core idea of the quasi-dependence account is to combine two criteria for causation: counterfactual dependence and intrinsicness. However, sometimes counterfactual dependence is an extrinsic matter as, for example, in cases of double prevention:

> "Suzy is piloting a bomber on a mission to blow up an enemy target, and Billy is piloting a fighter as her lone escort. Along comes an enemy fighter plane, piloted by Enemy. Sharp-eyed Billy spots Enemy, zooms in, pulls the trigger, and Enemy's plane goes down in flames. Suzy's mission is undisturbed, and the bombing takes place as planned. If Billy hadn't pulled the trigger, Enemy would have eluded him and shot down Suzy, and the bombing would not have happened" (Hall, 2004).

Double prevention means that a cause *c* prevents an event *d* that otherwise would have prevented the effect *e*. In this particular case Billy's pulling the trigger prevents Enemy's attack on Suzy. Otherwise Enemy's attack would have prevented the bombing of the target. The bombing of the target depends upon Billy's pulling the trigger. Yet this counterfactual dependence is not an intrinsic matter because it depends on the extrinsic absence of disabling factors. Let me explain. We shall call the sequence of Billy's attacking Enemy and the resulting bombing of the target by Suzy a process *S*. Consider a scenario that involves a duplicate of *S* but also a bomb placed under Enemy's seat which would have killed enemy just after Billy's attack on Enemy. Due to the bomb Suzy's destroying the target no longer depends upon Billy. Thus, the bomb is a factor that is extrinsic to *S* but undermines counterfactual dependence. Alternatively, the presence of the bomb would have to be counted as intrinsic. But that would lead to the implausible result that the bomb is a cause of the target's being destroyed as well (see (Hall, 2004, 245)).[6]

[6] This example has a structure that is similar to preemptive prevention cases that will be discussed

The conflict between intrinsicness and dependence has led Hall (2004) to the conclusion that we need to distinguish two concepts of causation: one, called 'dependence' that reflects intuitions associated with counterfactual dependence (and the intuition that omissions can be causes) and another, called 'production' that reflects intuitions associated with intrinsicness (alongside with locality and transitivity). We shall not discuss this proposal here in detail. What matters for our purposes is that due to the conflict between dependence and intrinsicness, quasi-dependence does not solve the problems associated with redundancy.

2.4.4. Causation as Influence

In his last attempt to account for redundancy Lewis goes back to fragility. Given that there are no clear criteria for event individuation, according to Lewis, we should make sure that our account of causation does not depend on this issue. For this reason Lewis introduces the notion of an alteration:

> "Let an *alteration* of an event E be either a very fragile version of E or else a very fragile alternative event which may be similar to E, but is numerically different from E" (2004, 88).

Going back to the Billy-Suzy case, one alteration of the actual bottle's shattering S is the shattering as caused by Suzy's stone (S_1).[7] This alteration is numerically identical to S and is instantiated. Another alteration of S is the shattering as it would have been caused through Billy's stone (S_2). This alteration is numerically different from S (and S_1) and not instantiated in the given scenario. The notion of alteration enables us to sideline questions regarding fragility. If we suppose that S is fragile, then S_1 and S_2 are alternative events. If we suppose that S is robust, then S_1 and S_2

in Chapter 6.

[7] Individuating the alteration with reference to Suzy threatens to make the account circular. It is close to trivial that Suzy's shattering the bottle depends upon Suzy's throwing her stone. But one could just as well describe S_1 by the exact position and velocity of the shattering pieces of glass.

are not alternative events. As long as we talk about alterations we do not have to decide this question.

The notion of an alteration can also be applied to the respective cause events. The supposed cause event is Suzy's throwing her stone ST. Instead we shall consider the numerically identical alteration ST_1 that specifies the exact time at which Suzy throws her stone. If Suzy had thrown a little earlier (alteration ST_2) than she actually did, this would have had an influence on the bottle's shattering to the effect that it also would have occurred a little earlier. The idea of Lewis's influence account of actual causation is to exploit this kind of dependence. More specifically, Lewis defines influence as a relation between a range of alterations:

> "C *influences* E iff there is a substantial range $C_1, C_2, \ldots$ of different not-too-distant alterations of C (including the actual alteration of C) and there is a range $E_1, E_2, \ldots$ of alterations of E, at least some of which differ such that if C_1 had occurred, E_1 would have occurred, and if C_2 had occurred E_2 would have occurred, and so on" (2004, 91).

In analogy to the earlier accounts, causation is defined as the ancestral of influence: "C causes E iff there is a chain of stepwise influence from C to E" (ibid).

Does the influence account solve the problems raised by the four kinds of of redundancy? Cases of symmetrical overdetermination are straightforwardly solved: for each of the two (or more) overdetermined causes there is a substantial range of alterations, such that alterations of the effect event depend counterfactually on them. For example, the forest fire would have happened earlier and in a different way if either one of the lightning bolts had occurred earlier and in a different way.

Next, consider late preemption. There is a substantial range of alterations of Suzy's throwing that will have an influence on the exact way the bottle shatters. However, also alterations of Billy's actions will have an influence on the exact way the bottle shatters. This was the reason why fragility was rejected as a solution to redundancy.[8] Lewis's response is that the notion of influence is gradual: there can be more or less influence. Consequently, according to Lewis, also the notion of cause is gradual: "Suzy's throw is much more of a cause of the bottle's shattering than Billy's" and this is because, "altering Suzy's throw while holding Billy's fixed would make a lot of difference to the shattering, whereas altering Billy's throw while holding Suzy's fixed would not" (Lewis, 2004, 92).

According to Lewis, one major advantage of the influence account is that it also accounts for trumping. Consider alternative orders that the major could have given to the corporal. These form a range of alterations and the corporal would act differently on receiving these orders. By contrast, the sergeant does not exert this kind of influence on the corporal. If the sergeant had given a different order, the corporal would still have obeyed the major and charged.

Let us take stock. So far I have presented four kinds of redundancy: symmetrical overdetermination, early preemption, late preemption, and trumping. These raise challenges for any account of actual causation and in particular those that are framed in terms of counterfactual conditionals. I have then discussed four kinds of counterfactual approaches to redundancy and have highlighted the difficulties that these accounts face. First, fragility is problematic because it leads to a profusion of causes. In particular, applying a fine-grained individuation scheme to the effect event does not solve the problem posed by preemption cases because it has the result that both preempting and preempted factors are identified as actual causes. Second, ancestral dependence fails with regard to late preemption (and symmetrical overdetermination and trumping) because in these cases there are no intermediate events that could form a chain of counterfactual dependence. Third, quasi-dependence faces problems because it involves a criterion of intrinsicness that is conflict with counterfactual dependence. Finally, the influence approach seems to account for

[8] This problem affects early preemption as well. There are tiny influences that the behaviour of the preempted backup assassin has on the exact way the victim is killed.

all kinds of redundancy. But like the fragility account it has the consequence that in early and late preemption both the preempting and the preempted factors are actual causes (but to a lesser degree than the preempting factor). Moreover, the influence account accounts for the asymmetry between trumping and trumped factors in cases like the one involving the major, the sergeant, and the corporal. This is an advantage over the other counterfactual accounts. Yet, as we will see in Chapter 4 this is still problematic because there is an important intuition according to which in some (to be specified) sense both the trumping and the trumped factors are actual causes. Finally, another potential problem for the ancestral dependence, the quasi-dependence, and the influence account is that they require actual causation to be transitive. We shall now turn to the problem of selection.

2.5. Causal Selection: Mill's Challenge

The problem of causal selection was first discussed by John Stuart Mill (1843/1882) in the context of his account of induction. As the "main pillar of inductive science" Mill considers the "Law of Causation" that "is but the familiar truth, that invariability of succession is found by observation to obtain between every fact in nature and some other fact which has preceded it" (236). The invariability of succession, however, seems to be violated in many cases. For example, suppose a person dies of eating a poisoned dish. According to Mill, this causal relation is grounded in the fact that death invariably succeeds eating poisoned food. However, there does not seem to be such an invariable sequence because there are instances where poisoned food is eaten but death does not follow, for example, because the person who ingested the food gets her stomach pumped.

Yet, according to Mill, this kind of counterexample does not threaten the invariability of succession. He argues that selective features of a situation such as eating a poisoned dish are not legitimately called a cause. Instead "[t]he real Cause, is the whole of these antecedents ; and we have, philosophically speaking, no right to give the name cause to one of them, exclusively of the others" (237). According to Mill, the invariability of succession is saved because once we take the whole of the antecedents into consideration the effects are determined. For example, eating the poisoned dish in conjunction with the fact that the stomach is not being pumped and the absence of all other potentially preventing factors is the total cause of the person's dying. If all these elements of the total cause are in place, then death invariably follows.

Mill justifies rejecting causal selection by arguing that "[n]othing can better show the absence of any scientific ground for the distinction between the cause of a phenomenon and its conditions, than the capricious manner in which we select from among the conditions that which we choose to denominate the cause" (238). By describing causal selection as capricious, Mill argues that there are no underlying criteria and *a fortiori* no rational criteria. Mill draws the revisionary conclusion that one should not perform causal selection and reserve the term cause for the totality of factors that could make a difference to the effect.

In the following I shall refer to Mill's argument against causal selection as *Mill's challenge*. In order to meet Mill's challenge we would, first, need to identify selection criteria that give rise to a descriptively adequate account of causal selection. Second, we would have to show that the criteria have rational grounds. This amounts to giving an explanation why we are justified to apply the respective rules of selection. These two points are different because, in principle, it might be possible to identify regularities in selective causal reasoning even if these regularities are not justified (they could simply be systematic biases).

Isn't the term "causal selection" suggesting an implausible description of causal reasoning? Usually we do not begin with a set of conditions sufficient for the effect and then select from this set the factors that appear most salient (Collingwood, 1938). More plausibly we come up with certain particular candidate factors and examine whether they are part of the total cause. Richard Wright (1985), for example, claims that this is the case in legal inquiry. If some harm occurred, then typically we do not ask for a set of sufficient conditions. Instead we begin by determining tortious behaviour and then investigate whether it is part of such a set of conditions. But even if "selection" is not an adequate description of the underlying *process* of causal reasoning, it points to an important challenge: why do certain factors have a privileged status as opposed to other factors that are considered to be mere background conditions?

Many philosophers follow Mill. For example, Lewis sidelines the issue of causal selection, arguing that he has "nothing to say about these principles of invidious discrimination" (Lewis, 1973a, 162).[9] Likewise, Hall argues that in causal selection "we typically make what are, from the present perspective, invidious distinctions, ignoring perfectly good causes because they are not sufficiently salient" (Hall, 2004, 228). Hall contrasts this with an "egalitarian sense of 'cause'" according to which "the complete inventory of a fire's causes must include the presence of oxygen and of dry wood" (ibid.).

In a sense, Lewis's and Hall's sidelining the problem of selection should be surprising, given that they take the problem of redundancy so seriously. Mill's argument against causal selection has a close analogy to fragility-based accounts of redundancy. Both accounts seem to explain away the respective problems by proposing a highly revisionary understanding of actual causation. In the case of fragility this involves an extremely detailed description of the involved events. As a consequence we would have to identify many more factors as actual causes than is typically acknowledged. In Mill's case this involves an exhaustive description of the

[9] But he addresses issues related to selective causal claims in his account of causal explanation (Lewis (1986b)).

conditions that precede the effect. The resulting notion of causation is much more restrictive than typically acknowledged. Thus, if fragility leads to an unsatisfactory account of redundancy, shouldn't we think of Mill's account as an unsatisfactory approach to selection?

2.6. Contextual Accounts

One reason for the popularity of Mill's view is that it is difficult to find definite and substantial criteria for distinguishing salient causes from background conditions. One such account is addressed by Mill. According to this account, we tend to identify events, that is, "instantaneous changes or successions of instantaneous changes" (237) as causes and identify states as background conditions. States, according to Mill, "possess [...] more or less of permanency ; and might therefore have preceded the effect by an indefinite length of duration, for want of the event which was requisite to complete the concurrence of conditions" (ibid.). Mill argues against this suggestion that the event "has really no closer relation to the effect than any of the other has" (ibid.). And, indeed, it is easy to find counterexamples. Suppose there is an explosion in Jones's flat because there was a gas leak. The instantaneous change that led to the explosion is Jones's lighting his cigarette but typically we also identify the state of gas being present as a salient cause of the explosion.

Another substantial account is suggested by Ducasse, who argues that "if a given particular event is regarded as having been *sufficient* to the occurrence of another, it is said to have been its *cause*; if regarded as having been *necessary to* the occurrence of another, it is said to have been a *condition of* it" (Ducasse, 1926, 58). But this does not seem to work better even in the most mundane cases of selection such as, for instance, with regard to the lightning that causes a forest fire under the condition that there is oxygen. The lightning is clearly not sufficient for the fire.[10]

But there is an alternative to accounts that give substantial criteria: contextual accounts. According to proponents of these accounts, the capriciousness of causal selection can be accounted for by the capriciousness of the contexts in which causal selection takes place. The earliest and most prominent contextual account is provided by Mackie (1974). According to Mackie, there are two reasons why we distinguish between causes and conditions. The first reason is associated with the *semantics* of causal claims and concerns what Mackie calls the causal field. According to Mackie, causal questions, such as "what caused this explosion?", can be expanded such that they seek explanation not for an event but to an "event-in-a-certain-field" (35). For example, the question for the cause of the explosion can be expanded to: "'What made the difference between those times, or those cases, within a certain range, in which no such explosion occurred, and this case in which an explosion did occur?'" The expansion can refer to the field as being "*this block of flats as normally used and lived in*" (35, emphasis original). One aspect of this causal field is that people in this block of flats normally strike matches to light their cigarettes. Thus, Jones's striking a match to light his cigarette is ruled out as an answer to the expanded question. It does not indicate a difference between the flat where the explosion occurred and other flats in the block where no explosion occurred. By contrast, gas leaks do not normally occur in this block of flats. So, the gas leak does not belong to the causal field and, thus, qualifies as a cause in the narrow sense.

The second reason why we distinguish between causes and conditions, according to Mackie, is associated with the pragmatics of causal claims. Mackie argues that "among factors not [ruled out by the causal field] we still show some preference" (35).

[10] Other suggestions refer to factors that we can control at all (Collingwood, 1938), the "event [that] is subject much more readily to operational control than others" (Beck, 1953, 374), or the factors that are particularly uncontrollable because they are volatile (Nagel, 1953, 698); for further examples and references see the discussion in van Fraassen (1980).

While the point about the causal field concerns the *meaning* of causal claims, Mackie considers these more specific preferences to be rather a conversational matter. We do not mention certain determinants because they happen to be irrelevant or simply less relevant than other determinants that earn the status of causes in the narrow sense.

Another kind of contextual account arises from approaches to causation as a contrastive notion. Causation as a contrastive notion is not a two-place relation as in 'C causes E' but a three place relation as in 'C causes E rather than E^{*}' or a four place relation as in 'C rather than C^{*} causes E rather than E^{*}.'[11] Schaffer (2005) argues that causal selection can be accounted for if we consider that the context of a causal claim influences the contrasts C^{*} and E^{*}. Thus, causal claims appear to be capricious because relevant parts of their semantics are kept implicit when the claims take binary form. Once we make the contrasts explicit capriciousness is eliminated.

Unlike Mackie, Schaffer (2012) considers selection to be exclusively linked to semantics and not to pragmatics. He gives three arguments. First, according to Schaffer, failures to select the suitable causes do not render a causal claim irrelevant but rather seem to render the claim false. For example, saying that the forest fire was caused by the presence of oxygen, according to Schaffer, is a claim that is not just an irrelevant truth, but false. Second, ordinary speakers assert the negation of claims that fail to select the suitable causes. For example, ordinary speakers assert sentences like "[t]he presence of oxygen did not cause there to be a forest fire, what caused the fire was the lightning" (43). Third, Schaffer argues that cancellation[12]

[11] Contrastive accounts of causation have been defended by Hitchcock (1996), Schaffer (2005), and Northcott (2008).

[12] Cancellation is a standard test to distinguish conversational implicatures from semantic entailments. Conversational implicatures can be blocked by an explicit negation. Schaffer gives the following example: "if I say of a job candidate that she has excellent handwriting, I can block the implicature that she is a poor philosopher by saying "but I don't mean to suggest that she is a poor philosopher" (2012, 44). By contrast, semantic entailments cannot be cancelled.

does not help for claims that fail to select the suitable causes. That is, according to Schaffer, we do not make claims like: "[t]he presence of oxygen caused there to be a forest fire, but I don't mean to suggest that the lightning strike played no role" (44).

So, according to both Mackie's and Schaffer's accounts, the binary causal expression 'C causes E' does not reveal the full meaning of a causal claim. The full meaning is captured if we fix an additional variable that reflects the relevant context. In Mackie's account the contextual variable is the causal field and in Schaffer's account it is the contrast of the cause and the contrast of the effect. Further contextual-variable accounts have been provided by Waters (2007) and Strevens (2011). Waters gives a contextual account that addresses the actual difference-makers in genetics. The contextual variable is the *population* (e.g. of fruit flies) that the geneticist looks at. According to Strevens, the contextual variable is the *framework*, a concept similar to Mackie's causal field.

Do contextual-variable accounts meet Mill's challenge? According to these accounts, the issue of causal selection arises only because the content of the contextual variable is left implicit. In a world of ideal causal speakers where the contextual variables are explicit the problem of causal selection would not even occur. Thus, if we want to understand contextual-variable accounts as providing a reply to Mill's challenge, then the proponents of these accounts argue that Mill's challenge results from a misunderstanding of the content of causal claims. The challenge is met if we use the causal idiom more carefully making sure that the contextual variable is made explicit.

But an important question remains unanswered: given a causal claim in binary form, what are the criteria for recovering the content of the contextual variable? Mackie does not tell us how to pick the relevant causal field. Schaffer states explicitly that "there is no obvious general procedure to recover the specific contrast applicable to the cause, or to recover the specific contrast applicable to the effect" (Schaffer, 2012, 36).

However, we often communicate successfully about causes in the binary form. Why is that the case? The explanation available for the contextualist is that we communicate successfully with binary causal statements because we share sufficiently similar contexts. Analogously, communication fails if we do not have sufficiently similar contexts. But under what circumstances are contexts sufficiently similar? In the following section we will see that there can be said more about this.

2.7. Normality

2.7.1. Hart and Honoré

Normality was introduced as a criterion for causal selection by Hart and Honoré (1959). Hart and Honoré's aim is an account of causation in the law that is based on a theory of common sense causal reasoning. The first part of their book *Causation in the Law* provides such a theory. Hart and Honoré's central concern is to identify among the factors that are necessary for an effect those factors that count as causes in the context of legal responsibility. Hart and Honoré describe two kinds of constraints. First, the factor must be either a voluntary action or an abnormal condition. Second, the factor should not be defeated by an event that occurs after the factor and before the effect. In the following I shall focus on the the first constraint. A more detailed discussion of the second constraint (which concerns the notion of proximate cause) will be provided in Chapter 6.

A voluntary action, according to Hart and Honoré, is one that is not "done 'unintentionally' (i.e. by mistake or by accident); or 'involuntarily' (i.e. where normal muscular control is absent); 'unconsciously', or under various types of pressure" such as those exerted by other persons, obligation or the lack of alternatives (38). The criterion of voluntary action, according to Hart and Honoré, puts constraints on the transitivity of causation. When we ask for the causes of some event how far do we have to trace back the causal history before we stop? Hart and Honoré argue that "[w]e do not trace the cause *through* the deliberate act" (40).

Abnormality and normality can be understood in two different ways, according to Hart and Honoré. First, functional normality describes the "usual state or mode of operation" of some object of inquiry. For example, if we investigate what caused a train accident, then the "normal speed and load and weight of [a] train and the routine stopping or acceleration" (32) are aspects of the usual mode of operation. These aspects are typically backgrounded because they do not not "'make the difference' between the accident and things going on as usual" (33). Second, normality can be related to duties or responsibility. For example, we identify the gardener's failure to water the flowers as a cause of the flowers' dying but not everybody else's failure because the gardener rather than everybody else is responsible for watering the flowers (see 35f).

Moreover, what counts as normal, according to Hart and Honoré, depends in two kinds of ways on the context. First, there is a contextual dependence on the kind of effect. Typically we do not identify oxygen as a cause of fire, but merely as a background condition. However, "[i]f a fire breaks out in a laboratory or in a factory, where special precautions are taken to exclude oxygen during part of an experiment or manufacturing process, since the success of this depends on safety from fire, there would be no absurdity at all in *such* a case in saying that the presence of oxygen was the cause of the fire" (33). The reason, according to Hart and Honoré, is that it is part of the normal functioning of the laboratory or factory processes that oxygen is absent. Thus, oxygen is not the cause of a fire in the woods, for example, but it is the cause of a fire in this kind of laboratory or factory. What counts as a cause of the fire depends on the particular kind of context where the fire occurs.

Second, there is contextual dependence on the interest of the causal reasoner.

This concerns examples, where "in one and the same case [...] the distinction between cause and conditions may be drawn in different ways" (33). For instance, [t]he cause of a great famine in India may be identified by the Indian peasant as the drought, but the World Food authority may identify the Indian government's failure to build up reserves as the cause and the drought as a mere condition" (33). The peasant, presumably, takes the government's food policy to be a normal state of affairs. Therefore, the peasant backgrounds this factor and categorizes it as a mere condition. The World Food authority, by contrast, takes the occurrence of a drought as something that is normal in a country with the climate and geography of India and identifies the government's failure to build up reserves as abnormal if compared to other such countries.

An immediate problem for Hart and Honoré's account is that there are causes that are neither abnormal nor voluntary acts. Lipton, for example, objects that "our ordinary notion of causation must allow for the beliefs that fire burns, sunlight warms, water quenches thirst, and innumerable other causal truisms at the heart of our ordinary conception of the world" (Lipton, 1992, 134).

This worry can be understood in two ways. First, it can be understood as pointing to a limitation of Hart and Honoré's criterion—a limitation that can be overcome by providing an alternative criterion of delineation between causes and conditions. In the following section we will discuss an extended notion of normality, according to which, for instance, the stone's being cold is a normal state and it's being heated is an abnormality.

Second, the worry can be understood as a more principled criticism regarding any account along these lines. Even an extended notion of normality will draw a line between causes and conditions. But maybe we should not draw such a line. For even if a factor is only a background condition, it has a causal role to play. I will not address this worry in detail in this chapter. But later we shall see that Hart and

Honoré's criterion (and related criteria) should not be understood as a criterion for distinguishing between causes and non-causes. Instead it is rather to be understood as a criterion for distinguishing *among* causes.

2.7.2. Extending Normality

According to Hart and Honoré, normality is defined either by functional or by moral and legal norms. In the more recent philosophical literature, however, the notion of normality has been given an even broader meaning. Sarah McGrath, for example, suggests a notion of normality that is "highly abstract" in the sense that "it is normal for x to ϕ iff x is *supposed* to ϕ" (McGrath, 2005, 138). What x is supposed to do can depend on a wide range of different standards, such as "*artifactual* standards governing what things like alarm clocks are supposed to do [...], *biological* standards governing what organs like hearts are supposed to do, [...] [and] *physical* standards governing the natural world" (McGrath, 2005, 138f).

While McGrath's artifactual standards are similar to Hart and Honoré's functional norms, the biological and physical standards need further explanation. One way to understand physical standards is provided by Maudlin. According to Maudlin, the structure of Newton's laws provides a way of distinguishing normal or default behaviour from abnormal or non-default behaviour. Maudlin argues that Newton's first law, the "*law of inertia*, [...] specifies [...] how the motion of an object will progress if nothing acts on it." The first law, then can be understood as describing an object's default or normal behaviour. "The second law then specifies how the state of motion of an object will *change* if a force is put on it" (2004, 430, emphasis original). In other words, the second law describes the non-default or abnormal behaviour of the object. Thus, Maudlin suggests a distinction between default and non-default behaviour along the lines of the distinction between inertial and non-inertial motion.

Maudlin also extends this distinction to what he calls "quasi-Newtonian" laws. Quasi-Newtonian laws hold whenever there are "inertial laws that describe how some entities behave when nothing acts on them, and then there are laws of deviation that specify in what conditions, and in what ways, the behavior will deviate from the inertial behavior" (ibid., 431).[13] This applies, for example, to the human biology: "[t]he inertial state of a living body is, in our usual conception of things, to remain living: That is why coroners are supposed to find a "cause of death" to put on a death certificate" (ibid., 434). Maudlin seems to think of intertial or quasi-inertial behaviour as underwritten by natural laws or other regularities employed by science. But if we extend this idea to regularities that inform common sense, we may be able to account for Lipton's counterexamples that we encountered in the foregoing section. The quasi-inertial state of an object may be construed in such a way, for example, that it remains cold. Relative to this inertial state sunlight would count as a cause of the object's warming.

2.7.3. Empirical Accounts

So far I have discussed accounts that argue for normality by appealing to causal intuitions. There have been also a number of empirical studies that investigate the role of normality in causal reasoning more systematically. This research can be traced back to studies on the availability of counterfactuals performed by Tversky and Kahneman (1973).[14] A key scenario in the more recent literature[15] is the pen vignette:

> "The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens,

[13] A similar distinction between quasi-inertial and quasi-non-inertial processes is employed by Hüttemann's "disruptive concept of causation" (2013; forthcoming).

[14] See Hitchcock (2011) for a discussion of the relation between availability of counterfactuals and the role of norms in causal reasoning.

[15] See Rose and Danks (2012) for an overview.

> 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 has repeatedly emailed them reminders that only administrative assistants 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" (Knobe and Fraser, 2008, 443).

The pen vignette describes a situation where the combination of the actions of two agents leads to an undesired outcome. The only difference between the two actions is that the Professor's taking the pen violates the department's policy while the administrative assistant conforms with the policy. Test subjects were asked to indicate whether they agreed or disagreed with the following statements: (1) 'Professor Smith is the cause of the problem.' (2) 'The administrative assistant is the cause of the problem.' Knobe and Fraser found that test subjects "agreed with the statement that Professor Smith caused the problem and disagreed with the statement that the administrative assistant caused the problem" (2008, 443). From this Knobe and Fraser infer that "moral judgments actually play a direct role in the process by which causal judgments are generated" (ibid.).

Knobe and Fraser's study indicates that prescriptive norms affect causal reasoning. But what is the role of statistical norms? Statistical norms do not prescribe a certain behaviour but merely describe what is common behaviour. While prescriptive and descriptive norms are often congruent, this is not necessarily the case, as Professor Smith's behaviour illustrates. Justin Sytsma, Jonathan Livengood and David Rose (2012) argue that we need to distinguish two types of typicality in order to characterize the role of statistical norms. First, there is typicality that relates to population-level statistical norms which describe "how *people* generally behave in a given type of situation" (816). This amounts to claims about how members of faculty or administrative assistants typically behave. Second, there is typicality that relates to agent-level statistical norms which describe how the particular agent "herself generally behaves in [a given type of] situation" (ibid.). This amounts to claims about how a particular agent such as Professor Smith or the administrative assistant behaves.

Based on variations of the pen vignette Sytsma et al. provide evidence to the effect that ordinary causal judgments are insensitive to considerations of population-level statistical norms. That is, even if the vignette describes Professor Smith's behaviour as statistically typical for members of faculty while the administrator's behaviour was described as statistically atypical for administrative assistants, test subjects identify Professor Smith as the cause to the same degree just as in the original study.[16] One could think that prescriptive norms override the statistical norms. But Sytsma et al. show that even in the absence of prescriptive norms, information about the population-level statistical norms does not influence the causal attribution.

Moreover, Sytsma et al. provide evidence to the effect that ordinary causal judgments are sensitive to typicality rather than atypicality in the case of agent-level statistical norms. That is, if the pen vignette specifies that it is unusual for Professor Smith to take pens (her taking the pen was an exception from her usual behaviour), then Professor Smith is less identified as a cause of the pen shortage.[17] In Chapter 5 we will have a closer look at these results and, in particular, Sytsma et al.'s claim that these results support their view that selective causal judgements of this kind are related to considerations about responsibility.

[16] A similar interpretation can be given to results provided by Roxborough and Cumby (2009)

[17] One might think that this rating is related to the fact that Professor Smith's repeated violation of the pen policy accumulates: the more often she takes pens the more likely it is that because of her actions there are no more pens. But Kirfel and Lagnado (2017) provide evidence to the effect that agent-level atypicality is also relevant in kinds of situation where this kind of accumulation does not occur.

The pen vignette and related examples take a central place in the recent research of experimental philosophers and empirical psychologists. Nevertheless, the use of these studies has also attracted criticism. For example, Samland and Waldmann argue that pen-vignette like studies do not track the influence of norms on causal judgements but rather on judgements of accountability. The authors claim that causal queries are ambiguous. "[I]n the context of human action [the term 'cause'] can both refer to the question of whether a mechanism underlying a causal relation is present and to the question whether an agent can be held *accountable*" (2016, 165). When presented with cases like the one described in the pen vignette, test subjects form hypotheses about which of the two senses of causation are intended. Moreover, the pragmatic contextual factors of the pen vignette favour an understanding of the causal query in terms of accountability. This is because (1) the causal relations are trivial (the relation between the taking of pens and the resulting problem is obvious), (2) the cover story refers to norm violation, and (3) the query concerns agents as causal relata. Their studies provide evidence to the effect that changing the pragmatic contextual factors can eliminate the influence of norm violation on causal judgements.

Another kind of criticism concerns the scope of the conclusions supported by studies such as those based on the pen vignette. According to Danks, Rose, and Machery, the evidential basis provided by such experiments does not warrant the claim that "[n]ormative considerations (broadly construed) influence causal cognition (broadly construed) and are perhaps even constitutive of various cognitive processes involved in aspects of causal cognition" (Danks et al., 2014, 254f). More specifically, the authors object that studies based on scenarios like the pen vignette involve learning causal relations from description and, thus, only probe a "highly language-driven kind of causal reasoning" (ibid., 256). Danks et al. provide evidence to the effect that if test subjects need to infer the causal relations from the cover story, then moral norms have no effect on the judgements.

2.7.4. Normality: A Reply to Mill's Challenge?

Proponents of normality typically emphasize the volatile character of normality. As we have seen above, Hart and Honoré, for example, argue that which norms are relevant depends on the specific effect and on the specific interests of the involved causal reasoners. Halpern and Hitchcock, who also propose to incorporate a notion of normality into their theory of actual causation, likewise argue that the resulting notion of causation is "(i) subjective, (ii) socially constructed, (iii) value-laden, (iv) context-dependent, (v) and vague" (2015, 431).

On an abstract level we can distinguish two ways in which normality depends on the context. First, normality is often understood as referring to a range of different kinds of norms, involving functional, legal, moral, and statistical norms. These different dimensions can have different relevance in different contexts. When I fail to get up in time because my alarm clock did not work, the functional sense of normality is relevant. It is difficult, by contrast to apply a moral notion of normality to this case. If a murderer kills someone, the moral or legal sense of normality applies but the functional sense does not. This kind of context-dependence seems to be particularly problematic when different norms apply to the same case but pull in opposing directions as, for example, in "[m]ost people speed. If the posted speed limit is 55 miles per hour, is driving at 55mph normal for conforming to the law, or abnormal for violating the statistical expectation?" (Blanchard and Schaffer, 2017, 193).

Second, even with regard to one kind of norm there is contextual variation. What is against the law, that is, legally abnormal is relative to the legal system. What is socially normal, is relative to the society you live in. What is statistically normal depends on the population that you look at. Physical standards such as Maudlin's criterion of inertial behaviour seem to be an exception. For whether the dynamics of a system is inertial or not does not depend on the reference frame. However, with Maudlin's extension of the concept to quasi-Newtonian laws even this criterion seems to become context-dependent. Mathias Frisch (2014) describes a case where the (non-inertial) trajectory of a satellite orbiting Earth is taken to be the satellite's default trajectory. Suppose the satellite has a rocket propulsion that makes it move uniformly along a straight line. Relative to the satellite's orbiting around Earth the resulting inertial trajectory is not a default.

Does this mean that accounts based on normality fail to meet Mill's challenge? Schaffer, for example, argues that Hart and Honoré's account deals with certain standard cases of selection "but at the price of such vagueness that it seems merely verbally distinct form the" view that it causal selection is capricious (2005, 343). However, even though normality is context-dependent it is not completely capricious. First, what is the relevant kind of norm in a given context is not nearly as arbitrary as opponents of normality seem to assume. Take, for example, the speeding case. Suppose one of the speeding cars is involved in an accident. Suppose also that the accident would not have happened if the driver had conformed to the posted speed limit. So the driver's behaviour violated the legal norm. In a legal inquiry that is concerned with finding the causes of the accident this will be the relevant norm. Whether or not most people speed is not important (unless they were also involved in the accident).

The same holds for what counts as normal according to each of the individual dimensions of normality. What is descriptively normal may depend on the relevant reference population. But once the reference population is fixed, objective statistical measures of normality can be defined. It is a matter of controversy whether an objective definition is possible for prescriptive norms. Yet, they are still relatively robust across a wide range of contexts. Most agents will agree about whether a particular artefact is functional or not. There is also considerable cross-cultural agreement about certain basic moral norms such as those that forbid killing humans, stealing, and lying (see e.g. Walzer (1994)). The robustness of normality considerations is also aptly illustrated by one of the standard examples of causal selection: a forest fire of which we say that it is caused by a lightning under the background condition of oxygen. Let us assume that this delineation is based on the assumption that we think that the presence of oxygen in the atmosphere is normal, whereas a lightning is abnormal. Surely it is possible to construct a situation that makes oxygen look like a salient cause and, thus, undermines the underlying notion of normality. Putnam gives an example: "Imagine that Venusians land on earth and observe a forest fire. One of them says, '*I* know what caused that—the atmosphere of that darned planet is saturated with oxygen'" (1982, 150). But note that Putnam goes as far as invoking visiting Venusians in order to make this point. Moreover, even though there is contextual variation, the choice of a norm in any particular context is not arbitrary. An artefact fulfils functional norms if it serves its purpose. So once the purpose of a particular artefact is agreed upon, the functional norms are fixed. Often (but not always) there are also independent reasons for why a particular set of legal norms holds in a particular country. Finally, many think that also with regard to certain basic moral principles we can give rational foundations, for example, in a Kantian tradition.

Thus, even if normality does not give hard and fast criteria for causal selection, it shows that selection is not entirely capricious as stated by Mill. Moreover, it seems that normality provides an explanation for why we are justified in selecting certain factors as salient (this was the second part of Mill's challenge as defined in section 2.5). Selecting the morally and legally abnormal, for example, seems to be justified in contexts where we are interested in assigning responsibility. Selecting the statistically and functionally abnormal seems to be justified in contexts where we are interested in changing some undesired outcome by means of intervention (Hitchcock and Knobe (2009)). A much more detailed discussion of this functional aspect of causal selection will be provided in Chapter 5.

2.8. Conclusion

In this chapter I have presented two problems for counterfactual accounts of actual causation: the problem of redundancy and the problem of selection. I have discussed several approaches to the problem of redundancy and I have highlighted the challenges that these approaches face. This overview prepares the discussion of causal model accounts of actual causation that will be given in the following chapters. Moreover, I have argued *pace* Mill and more recent opponents of causal selection that selection is to be taken seriously and that an informative response can be given in terms of notions of normality. I have also pointed out an analogy between the problem of redundancy and the problem of selection. Both problems can be explained away by referring to fragility and the notion of total cause, respectively. However, the strategy of explaining away the problems in these ways disregards an important sense in which judgements of actual causation are contextual. First, whether two factors are competing for being causes of the same effect event, depends on how fine-grained the description of the effect event is. And whether a fine-grained description is acceptable, depends on the context. Second, whether a factor is identified as an actual cause or merely as a background condition, depends on the relevant notion of normality. And whether a particular notion of normality is relevant, depends on the context.

3. Causal Models for a Unified Account?

In this chapter I will review causal-model based approaches that aim at a unified concept of actual causation. More specifically, I will begin by addressing two kinds of accounts. First, there are accounts that spell out actual causation in terms of counterfactual dependence given that certain other variables are held fixed at their *actual* values (as suggested by Pearl (1998; 2000), Hitchcock (2001), and Halpern (2015)). These accounts provide a solution to preemption cases. With regard to the Backup case, for example, the idea is that holding fixed the fact that the supervisor does not intervene, we can restore a counterfactual dependence of the victim's death on the actions of the assassin in training. And, therefore, the assassin in training is an actual cause. We will see that this kind of approach is too restrictive in order to account for cases of symmetrical overdetermination. Here we need to employ a second kind of approach according to which actual causation amounts to counterfactual dependence given that we set other variables to *non-actual* values (as suggested by Halpern and Pearl (2005)). With regard to the lightning case, for example, the idea is that under the contingency that the first lightning does not occur, there is a counterfactual dependence of the fire on the second lightning. And, therefore, the second lightning is an actual cause of the fire (and for reasons of symmetry the same holds for the first lightning).

Thus, it seems like the approaches based on causal models provide successful accounts of the instances of redundancy that have been so difficult to handle from the perspective of other counterfactual approaches.[1] However, the definitions have been shown to be too undemanding. There are instances where these definitions identify certain factors as actual causes that are typically not perceived to be actual causes. For example, consider a case called "Bogus Prevention" (Hiddleston, 2005). An assassin threatens to poison a victim's drink. The victim's bodyguard pours a harmless antidote in the drink but the victim would have survived anyway because the assassin had a last minute change of heart and decides not to administer the poison after all. We typically do not identify the antidote as an actual cause of the victim's survival because there was no need to neutralize any poison. But given the non-actual contingency that the assassin administered the antidote, there would have been a counterfactual dependence of the victim's survival on the bodyguard's action. And, thus, the bodyguard's actions do qualify as actual cause, according to Halpern and Pearl's definition.

Moreover, we will see that the model of Bogus Prevention is structurally isomorphic to models of symmetrical overdetermination cases—meaning that substituting the variables of one case by the variables of the other case leads to identical models. This is troubling for approaches like those discussed so far. These approaches rely on structural criteria in order to discern factors that are actual causes from factors that are not actual causes. But the fact that the structurally isomorphic cases give rise to different judgements seems to suggest that causal judgements are determined by non-structural features of the cases. This has been called the Problem of Isomorphism (Menzies (2004, 2007, 2017); Hitchcock (2007b); Hall (2007); Halpern (2008); Halpern and Hitchcock (2015)).

As a reaction there have been suggestions to extend the formalism by a distinction between default and deviant values that supposedly reflects the non-structural differences. Default values are those values that variables are expected to take on

[1] An exception are instances of trumping. I will address these in Chapter 4.

or that variables should take on. Thus, the default/deviant distinction explicitly involves subjective, context-sensitive, and norm-laden considerations. There are several ways of employing the default/deviant distinction. Here we will focus on an account that employs a normality ordering over possible worlds as suggested by Halpern and Hitchcock (2015).

An important consequence of employing such a normality ordering over worlds is that it takes into account the normality or abnormality of the purported cause variable. This means that Halpern and Hitchcock's approach seems to allow only those variables as actual causes that take on a deviant value. We will see that, thus, Halpern and Pearl's (2005) definition of actual causation, combined with normality orderings à la Halpern and Hitchcock promises to provide a unified concept of actual causation that deals not only with the problem of redundancy but also with the problem of selection.

The chapter is structured as follows. In section 3.1 I will introduce the formal framework of causal models. In section 3.2 I will review some criteria for choosing appropriate causal models. In section 3.3 I will present a first set of definitions of actual causation within this framework. The common idea of these definitions is that actual causation amounts to counterfactual dependence under the assumption that certain variables are being held fixed at their actual value. In section 3.4 I will turn to another kind of definition, according to which actual causation is related to counterfactual dependence under the assumption that certain variables are set to non-actual values. In section 3.5 I will discuss how the problem of late preemption has been addressed with the help of these definitions. In section 3.6 I will introduce the Problem of Isomorphism. In section 3.7 I will review suggestions to introduce a distinction between default and deviant behaviour. In section 3.8 I will show how the distinction is supposed to solve the Problem of Isomorphism and I will raise some initial doubts regarding this solution, which will prepare a more detailed discussion of defaults in Chapter 7. Finally, in section 3.9 I will briefly illustrate how Halpern and Hitchcock's normality ordering over worlds promises to provide a response to the problem of selection.

3.1. Causal Models

The use of structural models for representing causal relations goes back to Sewall Wright's (1920; 1921) work on the genetics of guinea pigs in the 1920s. Subsequenty, the framework has been applied fruitfully in econometrics and a wide range of social sciences including sociological, psychological, and political theory (see the overview in Goldberger (1972)). The version of the causal model framework that has been applied most fruitfully in the recent philosophical literature on causation goes back to the pioneering work of Peter Spirtes, Clark Glymour, and Richard Scheines (1993) and Judea Pearl (2000).

A causal model $\mathcal{M}$ is defined as an ordered pair, $\langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is a set of variables and $\mathcal{E}$ a set of structural equations.[2] The variables in $\mathcal{V}$ have two or more possible values. These values represent potential states of affairs or events in the model's target system. A variable's taking on one of the values represents an actual state of affairs or event in the target system.

The set of variables $\mathcal{V}$ has two disjoint subsets: the set of exogenous variables and the set of endogenous variables. The values of exogenous variables are determined by factors external to the causal model. They do not depend on the values of other variables in $\mathcal{V}$. The corresponding structural equations (sometimes called exogenous equations) simply ascribe a particular value to the exogenous variables. I will follow Halpern and Pearl's (2005) convention and summarize the content of exogenous equations by the *context* $\vec{u}$, a vector that specifies the actual value of each exogenous variable. The values of endogenous variables depend on the values of

[2] Henceforth I follow the exposition of the framework in Hitchcock 2001, if not stated otherwise.

other variables in $\mathcal{V}$. The corresponding structural equations (sometimes called endogenous equations) refer to exogenous and sometimes other endogenous variables on the right hand side. In deterministic causal models the values of endogenous variables are fully determined once the values of the exogenous variables are given.

The structural equations $\mathcal{E}$ provide a summary of the counterfactual dependencies that hold between the states of affairs or events. In a causal model there are as many structural equations as there are variables such that every variable appears on the left hand side of exactly one equation. Structural equations are *minimal*, that is, they must not refer to variables that the variable on the left hand side does not depend on. So if for all x', x, y, z, $Z = f_Z(x, y) = f_Z(x', y)$, then Z does not depend on X and X should be eliminated from f_Z. Moreover, structural equations are *complete* in the sense that they have to refer to all variables in $\mathcal{V}$ that the variable on the left hand side depends on. That is, if for some x, x', y, z, $f_Z(x, y) \neq f_Z(x', y)$, then Z depends on X and $f_Z(X, Y)$ is in $\mathcal{E}$.

Unlike mathematical equations, structural equations are not symmetric, so it matters whether a variable is on the left hand side or the right hand side of the equal sign. This reflects the asymmetry of causation. The syntax of structural equations is otherwise similar to the syntax of mathematical equations in that the value of the variable on the left hand side can be calculated by mathematical operations on the variables on the right hand side. The relevant operations include (but are not limited to) $\cdot, +, -$, as well as functions that select the minimum value and maximum value: $\min\{.,.\}, \max\{.,.\}$. Often the employed variables will be binary. Then it is useful to define structural equations in terms of symbols from sentential logic, which translate to mathematical equations as follows: $\neg X = 1 - X$; $X \vee Y = \max\{X, Y\}$; $X \wedge Y = \min\{X, Y\}$.

Models are sometimes defined as an ordered pair $\langle \mathcal{S}, \mathcal{E} \rangle$ (see e.g. Halpern and Pearl (2005)). $\mathcal{E}$ is again a set of structural equations and $\mathcal{S}$ is the *signature*, which is a tuple $(\mathcal{U},\mathcal{V},\mathcal{R})$. According to this notation, $\mathcal{U}$ is the set of exogenous variables, $\mathcal{V}$ the set of endogenous variables, and $\mathcal{R}$ is a function that associates with every variable in $\mathcal{U}$ and $\mathcal{V}$ a set of possible values. Unless stated otherwise, I will use the simpler notation where a model is denoted by $\langle\mathcal{V},\mathcal{E}\rangle$—with $\mathcal{V}$ referring to all variables of the model.

Causal models can be represented by *directed acyclic graphs* (DAGs). These are extremely useful for a quick (but incomplete) grasp of the causal relations of the target system. A causal graph is a set of nodes that correspond to the variables in $\mathcal{V}$. The nodes are connected by directed edges that correspond to the structural equations in $\mathcal{E}$. There is a directed edge leading from variable X to variable Y iff variable X features in the minimally complete structural equation for Y. In a causal graph exogenous variables are represented by nodes that have no incoming edges, whereas the nodes of endogenous variables always have incoming edges. Note that DAGs typically do not provide full causal information. The incoming arrows to the node of variable Y tell us what variables Y causally depends upon but they do not convey information about *how* Y depends on these variables. This information is provided by the structural equations.

It will be useful to introduce some terminology for family relations between variables. When a variable features on the right hand side of the equation for an endogenous variable, then it is a *parent* of this variable. Analogously, an endogenous variable is a *child* of every variable that features on the right hand side of its equation. Only endogenous variables have parents among the variables in $\mathcal{V}$. The notions of ancestor and descendant describe the transitive closure of the parent and child relation, respectively. So if A is a parent of B and B a parent of C, then A is an *ancestor* of C. Analogously, if C is a child of B and B is a child of A then C is a *descendant* of A.

It will also be useful to introduce the notion of a *directed path* or *route*. A directed path or a route is a sequence of variables $X_1, X_2, \ldots, X_n$ such that the X_i are parents of the X_{i+1} which means that a series of arrows leads from X_1 to X_n. Causal graphs are usually acyclic, that is, there are no directed paths or routes that lead from a variable X back to X. In other words acyclicity means that a variable is never its own ancestor or descendant.

3.2. The Art of Modelling

Given a particular target system, how do we arrive at a causal model of it? Unfortunately, there are no exhaustive criteria. In this sense generating a causal model is an art rather than a science (Halpern and Hitchcock (2010); Woodward (2016)). But the literature covers some pitfalls that are to be avoided plus some general rules of thumb.

Consider the selection of an appropriate set of variables $\mathcal{V}$. A causal model gives a partial representation of the target system. An important requirement is that the model "must include enough variables to capture the essential structure of the situation being modeled" (Hitchcock, 2007b, 503). This is of course only helpful to the extent that we have an idea of what the essential structure is. One way to think about this is to require that a causal model be stable in the sense that adding further variables to the model should not overturn causal judgements (Halpern and Hitchcock (2010); Halpern (2016)).

Next, consider the variables' possible values. First, the values of different variables should not represent events that stand in a logical relation. This is analogous to Lewis's (1986c) constraint that the events involved in causal claims be distinct. Suppose Martha says "hello" loudly. There is a counterfactual dependence of Martha's saying "hello" loudly on her saying "hello." But the counterfactual dependence is a logical dependence, not a causal dependence. Halpern and Hitchcock (2010) suggest to motivate this constraint as follows. Suppose we define one variable H_1 such that $H_1 = 1$ if Martha says "hello" and $H = 0$ otherwise and another variable H_2 such that $H_2 = 1$ if Martha says "hello" loudly and $H_2 = 0$ otherwise. The combination $H_1 = 0 \wedge H_2 = 1$ should be excluded because it is logically impossible for Martha to not say "hello" but say "hello" loudly.

Second, the values of any individual variable should be mutually exclusive (Halpern and Hitchcock, 2010). Suppose, for example, we want to express the dependence of the bottle's shattering on Suzy's throwing her stone. If $ST = 1$ represents the fact that Suzy throws her stone, then $ST = 0$ should imply that Suzy does not throw the stone. The requirement of mutual exclusivity is related to the contrastive character of causal claims, discussed in the foregoing chapter. If we claim that c rather than c^* causes e rather than e^*, then this implies that $c \neq c^*$ and $e \neq e^*$.

Another somewhat vague criterion is that the values of particular variables should not represent "unactualized possibilities we consider "too distant" to take seriously" (Hitchcock, 2001, 279). The rationale behind this requirement is that certain variables can easily become dependent upon other variables if we take scenarios into consideration that are very far fetched.

I conclude that the criteria for constructing apt causal models are vague and leave the concrete modelling decisions underdetermined. This is not necessarily a problem but is is an aspect of causal models that we shall keep in mind for the following discussion.

3.3. Early Preemption

In this section I shall address causal model accounts of actual causation that are motivated by the problems raised by cases that involve early preemption. The core idea of these accounts is that counterfactual dependence of the effect on the cause is being restored by fixing variables other than the cause and the effect variable at their *actual* values. Accounts along these lines have been suggested by Pearl (1998; 2000), Hitchcock (2001) and Halpern (2015). In this section I shall introduce in detail the two accounts by Hitchcock and Halpern.

Figure 3.1 displays a causal graph for "Backup," our example of early preemption. An assassin in training and her supervisor both set off to a mission to kill the victim. The assassin in training pulls the trigger of her gun ($T = 1$) and shoots the victim who dies instantly ($VD = 1$). If the trainee had not pulled the trigger, then her supervisor would have shot the victim instead ($S = \neg T$). The victim dies if either of the two assassins pulls the trigger of their guns ($VD = T \vee S$).

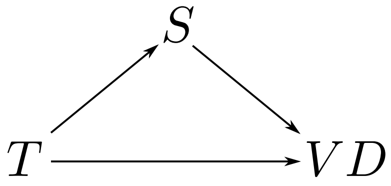

Figure 3.1.: Early preemption.

The trainee's actions are the actual cause of the victim's death. Yet VD's value does not depend upon T because the supervisor waits as a backup. Thus, straightforward counterfactual dependence fails as a criterion for actual causation. However, in the absence of the backup the counterfactual dependence is restored. That is, holding fixed the fact that the supervisor does not shoot (no matter what the trainee does), the victim's death does depend upon trainee's actions.

More generally the idea is that actual causation does not require straightforward counterfactual dependence. Instead it requires that there be certain variables such that if these variables are being kept fixed at their actual values, then the effect depends counterfactually on the cause. Judea Pearl (1998; 2000) was the first to make this idea explicit within the formal framework of causal models. According to Pearl, the defining criterion for $X = x$ to be an actual cause of $Y = y$ is that we can generate from the original causal model a reduced model (called a "natural beam") in which Y depends counterfactually on X. More specifically, we are required to arrive at such a reduced model by replacing the structural equations of variables other than X and Y by structural equations that simply set these variables to the values that they have in the actual context $\vec{u}$. The details of this particular account will not be relevant in the following. But we will see that Pearl's idea of generating a reduced causal model and then testing counterfactual dependence is common to all causal model accounts of actual causation that will be discussed here.

The idea of testing counterfactual dependence in a reduced causal model is central, for example, to Hitchcock's (2001) account of actual causation. Hitchcock motivates his account by considering the meaning of arrows in causal graphs. He takes them to express what he calls *explicitly non-foretracking counterfactuals*. On the common understanding counterfactuals are completely foretracking. Complete foretracking means, for example, that if T had taken on a different value, then S and VD would have taken different values as well, according to the structural equations. So, if we intervene on the trainee such that he does not shoot, the counterfactuals entail that the supervisor will shoot instead and that the victim will die anyway. The arrow from T to VD expresses a counterfactual that is not completely foretracking in the sense that it captures only the direct effect on VD while the effect on S is expressed by the other outgoing arrow.

More specifically, the arrow leading from T to VD means that there is *some* value s of S for which VD depends on T if S is held fixed at s, otherwise the minimality requirement would exclude T as an argument of the structural equation for VD. In "Backup" the value of S for which VD depends on T is the *actual value*. This is why the arrow from T to VD, according to Hitchcock, represents an *active* causal route, that is, a causal route along which causal influence propagates in the actual situation. Here is Hitchcock's definition of an active route:

> "Act: The route $\langle X, Y_1, \ldots, Y_n, Z\rangle$ is *active* in the causal model $\langle \mathcal{V}, \mathcal{E}\rangle$ if and only if Z depends counterfactually upon X within the new system of equations $\mathcal{E}'$ constructed from $\mathcal{E}$ as follows: for all $Y \in \mathcal{V}$, if Y is intermediate between X and Z, but does not belong to the route $\langle X, Y_1, \ldots, Y_n, Z\rangle$, then replace the equation for Y with a new equation that sets Y equal to its actual value in $\mathcal{E}$. (If there are no intermediate variables that do not belong to this route, then $\mathcal{E}'$ is just $\mathcal{E}$.)" (2001, 286).

The core idea, as in Pearl's account, is to test counterfactual dependence of the effect on the cause in a reduced causal model. This reduced causal model is generated from the original causal model by fixing those variables at their actual values that lie on a path from cause to effect other than the path under consideration. Hitchcock then identifies the existence of an active causal route between two variables as a necessary and sufficient condition for causation:

> "Let c and e be distinct occurrent events, and let X and Z be variables such that the values of X and Z represent alterations of c and e respectively. Then c is a cause of e if and only if there is an active causal route from X to Z in an appropriate causal model $\langle \mathcal{V}, \mathcal{E}\rangle$" (2001, 287).

Why is the process of the trainee's shooting the victim $\langle T, VD\rangle$ an active route, according to Hitchcock's definition? There is only one variable, S, that is intermediate between T and V but not on the route $\langle T, VD\rangle$. The actual value of this variable is $S = 0$, representing that the supervisor does not shoot. Thus, the new system of equations $\mathcal{E}'$ is: $T = 1$, $S = 0$, and $VD = T \vee S$. In this system of equations VD depends counterfactually on T. Thus, the trainee's pulling the trigger causes the victim's death.

Here is an example for an inactive route. Suppose the trainee fails to pull the

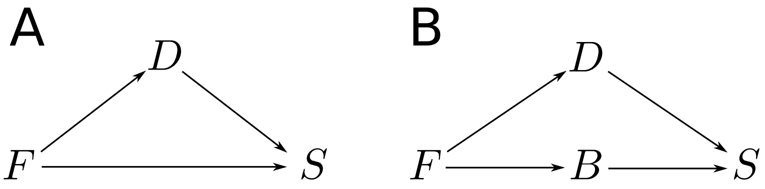

Figure 3.2.: Boulder.

trigger ($T = 0$). As a result the supervisor shoots the victim ($S = 1$ and $VD = 1$). Is $\langle T, S, VD \rangle$ an active route? No. Since there are no intermediate variables between T and VD that do not belong to this route, we have $\mathcal{E}' = \mathcal{E}$. But in $\mathcal{E}$ variable VD does not depend counterfactually on T. This reflects the intuitive judgement that the trainee's refraining is *not* a cause of the victim's death.

One problem for the accounts discussed in the previous chapter was that they employ assumptions about transitivity that lead to problems in cases like "Boulder." We shall now see how Hitchcock's causal route account deals with this kind of case. The discussion will also illustrate the model-dependence of Hitchcock's account.

Here is a causal model of "Boulder" (see figure 3.2A, for the DAG): $F = 1$ shall represent the boulder's falling, $D = F$ the fact that if the boulder were to fall, then Hiker would duck, and $S = \neg F \vee D$ the fact that Hiker would survive if either the boulder did not fall or Hiker ducked. According to Hitchcock's definition, the boulder's falling is not a cause of Hiker's surviving because there is no active route linking F and S. First, consider the direct route $\langle F, S \rangle$. The only variable that is intermediate between F and S is D. Replacing the structural equation for D by $D = 1$ and, thus, fixing D at its actual value does not render S counterfactually dependent on F. Second, consider the indirect route $\langle F, D, S \rangle$. There are no intermediate variables that do not belong to this route. Thus, the structural equations are not to be altered. Since S does not depend counterfactually on F in the original system of equations, the indirect route is inactive. Thus, there is no active route from F to S. Consequently, the boulder's falling is not an actual cause of Hiker's survival, which is the correct verdict.

Note that the route's $\langle F, D, S\rangle$ being inactive rests on the fact that there are no intermediate variables along the alternative route. This is not necessarily so. Suppose we interpolate a variable B between F and S, representing the boulder's approaching Hiker's head. This gives the following set of structural equations (see figure 3.2B, for the DAG): $F = 1$, $D = F$, $B = F$, and $S = \neg B \vee D$. Holding fixed $B = 1$ does render S counterfactually dependent upon F with the result that the boulder's falling *is* a cause of Hiker's survival, which is of course highly implausible.

Does the active route approach fail to account for Boulder after all? Hitchcock's strategy here is to reject the model with the interpolated variable because it is not an apt representation. An important requirement for apt representation, according to Hitchcock, is that the model should not represent scenarios that are not to be taken seriously. In order to see what this means, consider what kinds of scenarios are possible, according to this model. To begin with, take variables F and D as an illustration. Our including these variables in the model reflects that we take seriously two kinds of possibilities. First, there are possibilities that conform to the structural equations: either the boulder falls and Hiker ducks or the boulder does not fall and the Hiker does not duck. Second, there are possibilities that violate the structural equations. In particular, there is the possibility that the boulder falls but Hiker nevertheless does not duck. This is not unreasonable because we can imagine a situation where the boulder falls but Hiker fails to recognize the threat. Thus, F and D are unproblematic.

Let us now have a closer look at the interpolated variable B. In order to render $\langle F, D, S\rangle$ active as suggested above we have to make sure that the interpolated variable B does not lie on this path but only on the path $\langle F, B, S\rangle$. This is the case if B represents the presence of the boulder at a point where it is too late for Hiker to duck upon recognizing the threat. Again we need to check whether we are willing to take seriously all possible scenarios. But here is a problem. The new model requires us to imagine the scenario where the boulder approaches Hiker's head without having fallen. This is also the scenario that the active route criterion requires us to take into consideration: we need to check the counterfactual dependence of S on F holding fixed that the boulder approaches Hiker's head ($B = 1$). But this is, of course, implausible. Thus, including B into the model is to be rejected and the active route approach is saved.

Alternatively, suppose we live in a world where such mysterious appearances of boulders *are* a serious possibility. In this world hikers often die because boulders take them by surprise without having fallen. Suppose also that Hiker was lucky to encounter a boulder that is not of this mysterious kind such that it had to be dislodged before it could approach Hiker. In such a world it *is* plausible to think of the boulder's falling as a cause of Hiker's survival, which is again exactly the verdict reproduced by the active route approach.

I conclude that Hitchcock's active route criterion gives the correct result with regard to our example of early preemption and it does so without incorporating potentially problematic assumptions about transitivity. At the same time, the approach illustrates the relevance of the underlying modelling decisions. However, in the following section we will see that this definition faces problems with regard to other kinds of examples, especially with regard to cases of symmetrical overdetermination.

But before that we shall consider another, more recent, definition that works in terms of fixing variables at their actual values and that has been suggested by Halpern (2015; 2016). Here is the definition:

> $\vec{X} = \vec{x}$ is an actual cause of φ in $(M, \vec{u})$ iff
>
> AC1 $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$.

AC2$_\text{I}$ There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting $\vec{x}'$ of the variables in $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w}^*$, then

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}^*]\neg\varphi.$$

AC3 $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2$_\text{I}$.

Like Hitchcock's active route definition, Halpern's definition explicitly refers to a causal model M under a particular value assignment, which is given by the context $\vec{u}$. The first major difference is that Halpern's definition identifies an actual cause as a *set of variables* $\vec{X}$ whose elements X_i each take on a particular value x_i, as given by the vector $\vec{x}$. In Hitchcock's definition actual causes were defined as individual variables. I will address this difference in the following section.

The definition consists of three conditions. First, it has to be the case that both the cause event and the effect event have to be instantiated in the current model M and context $\vec{u}$. This is analogous to Hitchcock's account where the variables have to take on the values that represent the cause and effect event.

The second condition is a sophisticated counterfactual conditional. It requires that there be a set $\vec{W}$ of variables such that if we keep fixed these variables at their actual values $\vec{w}^*$, then changing the value of the cause variable will lead to a change in the effect.[3] Again, there is a similarity to Hitchcock's active route definition in that the variables in $\vec{W}$ are to be held fixed at their actual values.

However, there is also an important difference. Hitchcock's definition of actual causation depends on the activity of a single path linking cause and effect. The activity of the path is revealed by holding fixed intermediate variables on all other paths. Halpern's definition is more general in that it does not specify which variables have to be held fixed. This can be seen as an advantage for Halpern's definition because

[3] The sophisticated counterfactual reduces to a straightforward counterfactual if it is fulfilled with $\vec{W} = \emptyset$.

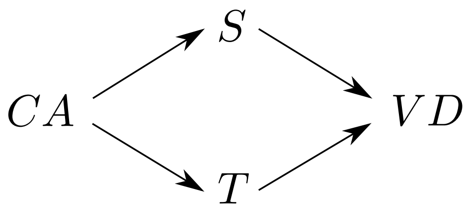

Figure 3.3.: The chief assassin's order is an actual cause of the victim's death.

it can handle cases like the following (see the DAG in figure 3.3, Halpern (2015)). Suppose the chief assassin orders two assassins (the trainee and his supervisor) to kill the victim. This time they are both supposed to shoot and each shot is lethal. That is, if the chief assassin issues her order ($CA = 1$) then both the trainee ($T = CA$) and the supervisor ($S = CA$) attack the victim and the victim dies ($VD = T \vee S$).

Presumably, the chief assassin's issuing her order is an actual cause of the victim's death. But, according to Hitchcock's definition, there is no active route linking the order to the victim's death. The route $\langle CA, T, VD\rangle$ is not active because if we keep fixed the actual fact that the supervisor attacks the victim, then the victim's death does not depend on the chief assassin's order. And the route $\langle CA, S, VD\rangle$ is inactive for the same reasons. Halpern's definition, by contrast, does identify the chief assassin as actual cause, with $\vec{W} = \emptyset$.[4]

Finally, Halpern's definition involves a minimality condition (AC3). AC3 guarantees that the definition does not count factors as causally relevant that are unnecessary for the effect. This condition does not have an analogue in Hitchcock's definition because Hitchcock's definition only considers individual variables.

In this section I have presented two definitions that make use of Pearl's idea that actual causation amounts to counterfactual dependence in a reduced model, where the reduced model is generated by fixing variables other than the cause and the effect

[4] There is another difference. In Hitchcock's definition the variables Y that are to be held fixed are identified as *intermediate* variables. Halpern's variables $\vec{W}$ do not have to be intermediate. This difference, however, is not important. Any variable that is causally relevant for the effect e but does not lie on one of the paths from X to Z will not be affected by an intervention on X and, thus, will keep its actual value without being held fixed.

at their actual values. These definitions are successful with regard to standard cases of early preemption. One crucial advantage of these definitions over alternative definitions discussed in the foregoing chapter is that, given an appropriate model, they do not fail with regard to cases such as "Boulder." But what about the other forms of redundancy?

3.4. Symmetrical Overdetermination

Unfortunately, the definitions of actual causation provided in the previous section do not apply to cases of symmetrical overdetermination. Suppose both the assassin in training ($T = 1$) and his supervisor ($S = 1$) shoot at the same time and that both their bullets hit the victim lethally ($VD = T \vee S$, see the graph in figure 3.4). In this kind of case we identify both the trainee's and the supervisor's actions as actual causes.

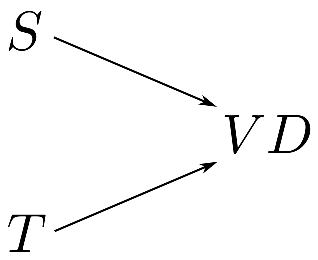

Figure 3.4.: Symmetrical overdetermination.

However, the definitions discussed so far do not not give this result. In the reduced causal model where the supervisor variable is kept fixed at its actual value $S = 1$ there is no counterfactual dependence of VD on T. Given that the supervisor pulls the trigger of her gun, the victim will die, no matter what the assassin in training does.

An exception is Halpern's definition. Here an actual cause can take the form of a conjunct of two variables. If we assume that $\vec{X} = \{S, T\}$, then there is a counterfactual dependence of the effect on $\vec{X}$. However, this definition does not reflect the fact that either of the two assassins is an individual actual cause of the victim's death. This is particularly problematic if we look at the case from an interventionist perspective: in order to save the victim we need to apply two interventions, one directed at the assassin in training and one directed at the supervising assassin. I will get back to this problem in Chapter 4.

There is a way to save the basic idea that actual causation amounts to counterfactual dependence under certain contingencies. The victim's death does depend on the trainee's attack in situations where the supervisor does not pull the trigger of her gun. That is, if we set the supervisor variable to the non-actual value $S = 0$, then VD depends counterfactually on T. In general, the idea is that in cases of symmetrical overdetermination, the counterfactual dependence between cause and effect can be restored if we set certain factors to *non-actual* values.

This kind of counterfactual dependence was already envisioned by Pearl's (1998; 2000) original theory, where he uses it in order to define a notion of contributing cause. The details of Pearl's definition shall not concern us here.[5] Instead we shall look at the more developed definition provided by Halpern and Pearl (2005). In order to introduce this definition let me discuss the following preliminary suggestion:[6]

AC2$_{IIpre}$ There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}']\neg\varphi.$$

The difference between AC2$_{IIpre}$ and AC2$_{I}$ is that AC2$_{IIpre}$ allows us to set the

[5] It is known to face counterexamples, as indicated below.

[6] In order to arrive at a definition of actual causation AC2$_{IIpre}$ has to be seen in combination with AC1 and AC3 as presented in the foregoing section. I will not repeat these conditions here and in the following.

variables in $\vec{W}$ to non-actual values $\vec{w}'$ whereas AC2$_I$ only allows us to keep them fixed at their actual values $\vec{w}^*$. With this amendment we get $T = 1$ as an actual cause of $VD = 1$ with $\vec{W} = S$ and $\vec{w}' = 0$, and analogously for $S = 1$.

At a first glance it also seems to be the case that this preliminary definition still gives the correct verdict with regard to early preemption. It still allows us to reduce the model of "Backup" to a model where $S = 0$ is fixed such that the trainee is classified as an actual cause of the victim's death. Moreover, it reproduces the judgement that in the current situation the supervisor's behaviour is not an actual cause of the victim's death—she does not even attack the victim. Formally, this can be seen as follows: in the given situation we have $(M, \vec{u}) \models S = 0 \wedge VD = 1$ (condition AC1). Then condition AC2$_{IIpre}$ cannot be fulfilled because if we set $S = 1$, there is no possibility for the victim to survive (that is, there is no possibility for $\neg\varphi$), which means that AC2$_{IIpre}$ is not fulfilled.

However, consider a model of "Backup" that includes variables that inform us about whether the assassins set off to their missions (see figure 3.5). In the actual situation both the assassin in training and her supervisor set off to a mission to kill the victim ($TM = 1$ and $SM = 1$). The assassin's pulling the trigger then depends on whether she sets off to the mission ($T = TM$) and the supervisor's pulling the trigger depends upon the supervisor's setting off and the trainee's pulling the trigger ($S = SM \wedge \neg T$).[7]

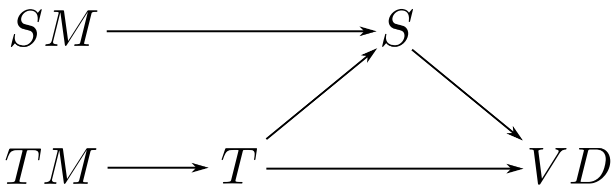

Figure 3.5.: An expanded model of "Backup."

[7] In the following I will use this expanded model of the early preemption case in order to motivate Halpern and Pearl's version of condition AC2. In their treatment the condition seems to be motivated by cases of late preemption instead. However, in section 3.5 we will see that their model of late preemption is problematic.

With regard to this model the amended definition gives the counterintuitive result that the supervisor's setting off to her mission is an actual cause of the victim's death. If we choose $\vec{W} = \{TM, T\}$ with $TM = 0$ and $T = 0$, then the value of VD depends on SM. This is problematic because this more complex causal model seems to be perfectly legitimate.

The problem can be avoided if we choose $\vec{W}$ more carefully. The problem is that by setting T to $w' = 0$ we establish a causal process leading from the supervisor's setting off to the mission $SM = 1$ via the supervisor's pulling the trigger of her gun $S = 1$ to the victim's death $VD = 1$. And this causal process was not there in the actual situation. This can be avoided by taking into account the processes that are active in the actual situation, as follows:

AC2$_{II}$ "There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that if $(M, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}'] \neg \varphi$ [...].

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*] \varphi$ for all subsets $\vec{W}'$ of $\vec{W}$[8] and all subsets $\vec{Z}'$ of $\vec{Z}$ [...]" (Halpern and Pearl, 2005, 853).

In the following I will refer to this as the HP definition of actual causation. In analogy to Hitchcock's account (in most cases[9]) the variables in $\vec{Z}$ can be understood as constituting an active causal route that connects the cause $\vec{X} = \vec{x}$ and the effect φ. The account differs from Hitchcock's account in that the off-path variables $(\vec{W})$ can be set to some non-actual value $\vec{w}'$. This is expressed by condition AC2$_{II}$a. But we have seen that condition AC2$_{II}$a is too permissive because it does not track whether the change in the effect from φ to $\neg\varphi$ is due to the change in $\vec{X}$ or due to the change

[8] An earlier definition by Halpern and Pearl (2001) as well as Pearl's definition of contributing cause in (Pearl, 1998, 2000) skips the requirement "for all subsets $\vec{W}'$ of $\vec{W}$." But this leads to problems, as shown by a counterexample provided by Hopkins and Pearl (2003).

[9] See Halpern (2016, 62ff) for a discussion of counterexamples.

in $\vec{W}$. With AC2$_{II}$b Halpern and Pearl add a restriction that picks out those changes in φ that are due to the change in $\vec{X}$. The condition states that setting any subset of variables in $\vec{W}$ to non-actual values should not affect φ given (1) that we keep the supposed actual cause fixed at its actual value $\vec{X} = \vec{x}$ and (2) that all the variables in an arbitrary subset of the variables along the active causal route are fixed at their actual value ($\vec{Z}' = \vec{z}^*$).

Let me illustrate how this amended condition helps with regard to the expanded model of early preemption. We shall begin with the trainee's setting off to the mission TM. AC1 is fulfilled because in the actual situation it is the case that the trainee sets off to the mission ($TM = 1$) and the victim dies ($VD = 1$, in the following abbreviated by φ). AC2$_{II}$ is fulfilled if we partition $\mathcal{V}$ such that $\vec{Z} = \{TM, T, VD\}$ and $\vec{W} = \{SM, S\}$. AC2$_{II}$a is fulfilled for this partition because the victim does not die ($\neg\varphi$) if neither the assassin in training nor the supervisor set off to their missions ($TM = 0$ and $SM = 0$). AC2$_{II}$b is fulfilled because given the fact that the assassin in training sets off to her mission, φ cannot be affected by changing SM or S, which are elements of $\vec{W}$.

Now turn to SM. The supervisor's setting off to the mission is not an actual cause of the victim's death because there is no partition such that AC2$_{II}$a and AC2$_{II}$b are fulfilled at the same time. The only partitions that render VD counterfactually dependent upon SM are (i) $\vec{Z} = \{SM, S, VD\}$, $\vec{W} = \{TM, T\}$ with $\vec{w}'$ such that $TM = 0$ and $T = 0$ and (ii) a partition with only $W = \{T\}$. But for these partitions condition AC2$_{II}$b is not fulfilled. There exists a subset $\vec{Z}' = \{S\}$ of $\vec{Z}$ such that if we fix $\vec{Z}' = z^*$ (that is: $S = 0$), then the effect does not occur ($\neg\varphi$). Thus, the definition reproduces our intuitive judgement that the trainee's actions are an actual cause of the victim's death but not the supervisor's actions.

Thus, the Halpern-Pearl definition of actual causation seems to be superior to Hitchcock's active route criterion because it handles cases of symmetrical overdetermination as well as cases of early preemption.

3.5. Late Preemption

In Chapter 2 we have seen that late preemption is particularly problematic for counterfactual accounts. Unlike early preemption it cannot be solved by identifying an intermediate event that, together with the presumed cause and effect event, would constitute a chain of counterfactual dependence. In this section I will discuss two models that have been employed by Halpern and Pearl (2005) in order to tackle late preemption. The first model is by far the most common model in the current literature but is problematic for a number of reasons. The second model is a time-indexed model that is much less common but more appropriate. The review of these approaches has two purposes. First, I shall assess the advancement that the causal model framework provides with regard to late preemption. I will argue that only the time-indexed model brings genuine progress. Second, this section shall prepare a discussion of late preemption in the context of intervention and responsibility that will be provided in Chapters 4 and 5.

Here is a common way to model the Suzy-Billy example of late preemption.[10] Let $ST = 1$ and $BT = 1$ represent that Suzy and Billy both throw stones towards the bottle, respectively. Moreover, let $SH = 1$ and $BS = 1$ represent Suzy's hitting the bottle and the bottle's shattering. Finally, $BH = 0$ shall represent Billy's not hitting the bottle. The causal structure shall be represented by the causal model displayed

[10] Among others this model is suggested by Halpern and Pearl (2001); Hopkins and Pearl (2003); Halpern and Pearl (2005); Hitchcock and Knobe (2009); Halpern and Hitchcock (2010); Halpern (2015, 2016); Menzies (2017); Hitchcock (2017).

in figure 3.6, with the following structural equations:

$$SH = ST,$$

$$BH = BT \wedge \neg SH,$$

$$BS = SH \vee BH.$$

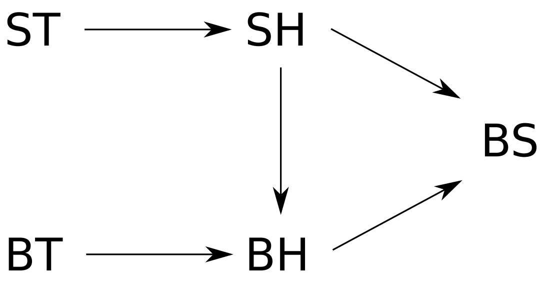

Figure 3.6.: Causal graph of late preemption: Billy and Suzy throw stones at a bottle. Suzy throws a little earlier, her stone hits the bottle and the bottle shatters. Billy's stone passes the initial position of the bottle only moments later. If Suzy had not hit the bottle, then Billy would have hit and shattered it.

With this model it should not come as a surprise that the definitions presented in the previous sections account for late preemption. For this model has exactly the same structure as the expanded model of early preemption discussed in the foregoing section. In order to see the structural equivalence, simply replace Suzy's throwing her stone and hitting the bottle by the trainee's setting off to the mission and his pulling the trigger of his gun. Likewise replace Billy's actions with those of the supervisor and the victim's death with the bottle's shattering.

However, there are two major problems with this model.[11] First, according to

[11] Another problem is that the relation between *ST* and *SH* seems to be at least partly conceptual because Suzy's hitting the bottle requires that Suzy has thrown her stone (the same holds for the relation between *BT* and *BH*). This means that the model violates one of the rules of apt modelling detailed in section 3.2. The problem is that this would allow a scenario where Suzy's stone hits the bottle without Suzy having thrown the stone. Yet, this is not a deep problem. We can redefine the variables. Suppose Suzy throws from the right side and Billy from the left. Accordingly, the intermediate variables can be defined as *RH* describing whether the bottle is hit from the right side and as *LH* whether the bottle is hit from the left side. If Suzy hits the bottle, the bottle will be hit from the right side and if Billy hits the bottle, the bottle will be hit from the left side. These

the model, *BS* is not a cause of *BH*. This is clearly wrong. Billy's not hitting the bottle is caused by the bottle's being broken at the time when Billy's stone arrives. In order to reflect this relation one would have to include a directed edge leading from *BS* to *BH*. Second, the model takes *SH* to be a direct cause of *BH*. But this is clearly wrong as well. Direct causation requires that there be an intervention on *SH* that changes the value of *BH* given that all other variables are kept fixed at some value. But this is not the case. For suppose we keep fixed the fact that the bottle is shattered ($BS = 1$, for example, because there is a third person who throws a stone and destroys the bottle after Suzy but before Billy). Then intervening on *SH* will not make a difference to the value of *BH*. Likewise, if we hold fixed that the bottle is not shattered, then intervening on Suzy's hitting will not make a difference to Billy's hitting the bottle. For instance, think of a situation where Suzy's stone does not have not enough momentum to destroy the bottle.

This is so troublesome because the model thereby *essentially* misrepresents late preemption. The difference between early and late preemption is that in late preemption the alternative process only fails because the effect has already occurred. That is, the only influence of the preempting process on the preempted process is via the effect event. This requirement is violated by the direct causal relation between *SH* and *BH*. In fact, by including this causal relation the Suzy-Billy case is turned into a case of early preemption, as the structural equivalence with the assassin case illustrates. Consequently, this treatment of late preemption does not genuinely advance the discussion if compared to Lewis's account of ancestral dependence (aside from the fact that this treatment does not require transitivity).

Luckily, however, Halpern and Pearl give also an alternative model (which is much less recognized) that represents the relation between *SH*, *BH*, and *BS* more appropriately (and also avoids the conceptual relation between, e.g., *ST* and *SH*).

relations are causal.

This model is governed by two kinds of equations that incorporate time indices:

$$\begin{aligned} H_i &= T_i \wedge \neg BS_{i-1} \\ BS_i &= BS_{i-1} \vee H_i. \end{aligned}$$

According to the first equation, the bottle is hit at time $\tau = i$ if someone throws the stone at $\tau = i$ and the bottle was not already shattered at time $\tau = i - 1$. According to the second equation, the bottle is shattered at time $\tau = i$ if it was shattered earlier ($\tau = i - 1$) or if it is hit at time $\tau = i$. Plugging in the details of the Suzy-Billy case yields the following set of structural equations:

$$\begin{aligned} H_1 &= ST, \\ BS_1 &= H_1, \\ H_2 &= BT \wedge \neg BS_1, \\ BS_2 &= BS_1 \vee H_2, \\ H_3 &= T_3 \wedge \neg BS_2, \\ BS_3 &= BS_2 \vee H_3. \end{aligned}$$

Here T_3 and H_3 are added in order to represent the time-invariance of the model. Figure 3.7 displays the corresponding causal graph.

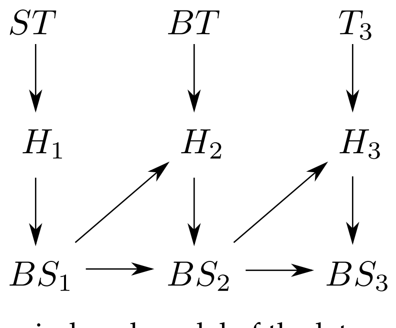

Figure 3.7.: Time-indexed model of the late preemption case.

The HP definition reproduces the correct causal judgments with regard to this model. First, $ST = 1$ is an actual cause of $BS_3 = 1$. Take $\vec{W} = \{BT\}$ and $\vec{w}' = 0$. Then setting $ST = 0$ leads to $BS_3 = 0$ (AC2$_{II}$a fulfilled). Moreover, if $ST = 1$ (Suzy throws), $BT = 0$, and keeping fixed any subset of $H_1 = 1$, $BS_1 = 1$, and $BS_2 = 1$ still gives $BS_3 = 1$ (AC2$_{II}$b fulfilled). By contrast, $BT = 1$ is not an actual cause of $BS_3 = 1$ because there is no partition $\vec{Z} \cup \vec{W}$ that fulfills both AC2$_{II}$a and AC2$_{II}$b. H_2 must be assigned to $\vec{Z}$ and BS_1 must be assigned to $\vec{W}$ with $\vec{w}' = 0$. Otherwise BS_3 does not depend counterfactually on BT (AC2$_{II}$a). But then there are subsets $\vec{W}'$ and $\vec{Z}'$ that violate AC2$_{II}$b. If we set $BS_1 = 0$ and $H_2 = 0$, then $BS_3 = 0$. Thus, Halpern and Pearl provide a model that avoids the problems of the model in figure 3.6 and where the HP definition still gives the correct verdict.

Note that the time-indexed model essentially distinguishes between different events of bottle shattering. According to this model, for example, $BS_1 = 1$ needs to represent an event that is distinct from $BS_2 = 1$. Thus, the model requires an extremely fine-grained scheme for event individuation.[12] In Chapter 2 we have seen that such a fine-grained individuation has been considered to be problematic. The main worry, according to Lewis, is that there are no principled guidelines for when such a fine-grained individuation scheme is appropriate and when it is not. However, this worry does not seem to apply to Halpern and Pearl's time-indexed model. There is a rationale behind choosing a fine-grained individuation scheme in this context. Only the fine-grained individuation allows us to capture an important feature of the causal structure of late preemption cases: Suzy's hitting the bottle causes it to shatter, the bottle's being shattered is a cause of Billy's failing to hit the bottle, which is the reason why Billy does not shatter the bottle at a later time.

[12] Another problem is that it would be incoherent to speak of $BS_1 = 1$ and $BS_2 = 1$ as representing events. The shattering of the bottle at $\tau = i$ cannot be a cause of the shattering of the bottle at time $\tau = i + 1$. If anything at all, the shattering at $\tau = i$ is a cause of the bottle's *not* shattering at $\tau = i + 1$. Instead, we should rather think of BS_1 and BS_2 as describing states. The bottle's being shattered at time $\tau = 1$ can be seen as a cause of its being shattered at $\tau = i + 1$. I shall not try to resolve issues associated with the notion of event here.

I conclude that the definitions of actual causation that use the causal model framework have made advancements with regard to cases of late preemption. Halpern and Pearl's time-indexed model reproduces the relevant causal intuitions by employing a fine-grained individuation scheme with regard to the effect event. In the context of late preemption the fine-grained individuation scheme is independently justified because it reveals essential aspects of the underlying causal structure. At the same time, the treatment of late preemption illustrates the relevance of the underlying causal model's being an apt model—a restriction that shall be particularly relevant in the following discussion.

3.6. The Problem of Isomorphism

The use of causal models has without doubt increased the degree of precision in discussions of redundancy cases. However, more recently philosophers of causation have argued that there are principled limitations for accounts of actual causation that are based on standard causal models. The pessimism arises from a range of cases where causal judgements supposedly cannot be captured by these standard causal models. The most prominent problem is the Problem of Isomorphism (Menzies (2004, 2007, 2017); Hitchcock (2007b); Hall (2007); Halpern (2008); Halpern and Hitchcock (2015)). There are examples of pairs of cases that appear to have isomorphic counterfactual structure, yet, different causal judgements apply. This is taken to indicate that our causal judgements depend on considerations that go beyond the counterfactual structure and, thus, beyond the content of standard causal models. Therefore, many have suggested extending causal models such that they distinguish between default and deviant behaviour. The idea is that through this extension the diverging causal judgements can be explained.

Instances of the Problem of Isomorphism involve pairs of cases that have isomorphic causal models. Two causal models are isomorphic iff their "patterns of counterfactual independence and dependence are identical, *modulo* the use of different variables" (Menzies, 2017, 159). That is, two causal models M_1 and M_2 are isomorphic iff the structural equations of model M_1 can be generated from the structural equations of M_2 by substitution of variables, and vice versa.

Before we continue let me make a remark about terminology. There is a question regarding what exactly is taken to be isomorphic. According to Menzies, the pairs of examples are isomorphic on the level of *counterfactual* structure and diverge with regard to *causal* structure (2017, 154). By contrast, Halpern and Hitchcock (2015) take the causal structure to be isomorphic and say that the cases diverge with regard to judgements of actual causation. Both terminological choices, however, are imprecise in the sense that they suggest an isomorphism between the target systems. However, what is taken to be isomorphic are only the models of the pairs of cases.[13] A detailed discussion of this point will be provided in Chapter 7. In the following I shall speak of isomorphisms as a relation between causal models. The relations within these models will be referred to as structural equations. From this I distinguish judgements or claims of actual causation. These judgements or claims may diverge in instances of the Problem of Isomorphism.

Here is an example of a case (called "Careful Poisoning") that is structurally isomorphic to "Backup":

> "Assistant Bodyguard puts a harmless antidote in Victim's coffee. Buddy then poisons the coffee, using a type of poison that is normally lethal, but which is countered by the antidote. Buddy would not have poisoned the coffee if Assistant had not administered the antidote first. Victim drinks the coffee and survives" (Hitchcock, 2007b, 519).

[13] Correspondingly, substitution of variables is an operation that is performed at the level of causal models, not at the level of target systems.

Here is a causal model of the case: first, A is a two-valued variable representing whether there is antidote in the coffee ($A = 1$) or not ($A = 0$). Second, P represents whether there is poison in the coffee ($P = 1$) or not ($P = 0$). Third, VS represents whether Victim survives ($VS = 1$) or not ($VS = 0$). The structural equations are as follows:

$$A = 1$$

$$P = A$$

$$VS = A \vee \neg P$$

The structural isomorphism with the standard causal model of "Backup" can be revealed by substituting A by T (the actions of the assassin in training), P by $\neg S$ (the negation of the supervising assassin's actions) and VS by VD (describing whether the victim dies)[14]—which gives exactly the structural equations presented in section 3.3.

The structural isomorphism entails that in both cases the same counterfactual dependencies and independencies hold. In Careful Poisoning the victim's survival does not depend on the bodyguard's administering the antidote. Likewise, in "Backup" the victim's survival does not depend upon the trainee's actions. This is because in both cases the influence is mediated by two causal paths that cancel out. Moreover, in both cases the effect depends on these factors if the third factor is held fixed at its actual value. If we hold fixed that Buddy administers the poison ($P = 1$), then the Victim's survival does depend on the bodyguard's actions ($VS = A$). Likewise, if we hold fixed that the supervisor does not shoot ($S = 0$), then the victim's death depends on the trainee's actions ($VD = T$).

In Careful Poisoning the victim's survival depends counterfactually on the body-

[14] Note the particular value assignment: in Careful Poisoning VS describes whether the victim *survives* (1) or not (0) and in Backup VD describes whether the victim *dies* (1) or not (0).

guard's action given that we hold fixed the fact that Buddy administers his poison. This means that the definitions provided by Hitchcock (2001), Halpern (2015), and Halpern and Pearl (2005) identify the bodyguard's administering antidote as an actual cause of the victim's survival. But this seems to be wrong. The bodyguard's administering antidote is the reason why Buddy poisons the coffee in the first place.

So "Careful Poisoning" is a counterexample to definitions like those introduced in sections 3.3 and 3.4. One way to respond to this counterexample would be to search for a refined definition that excludes the bodyguard's administering antidote as an actual cause. However, the problem is that any such refinement would need to exclude the trainee's actions in Backup as actual cause as well. The reason for this is that the bodyguard and the trainee are equivalent from the perspective of the causal models. This is why the two cases form an instance of the Problem of Isomorphism.

Here is another example of two cases that have isomorphic causal models but give rise to conflicting causal judgements. Consider "Bogus Prevention" (going back to Hiddleston (2005)).

> "Assassin is in possession of a lethal poison, but has a last-minute change of heart and refrains from putting it in victim's coffee. Bodyguard puts antidote in the coffee, which would have neutralized the poison had there been any. Victim drinks the coffee and survives" (Halpern and Hitchcock, 2015, 428).

The case can be represented with the following causal model. First, let $P = 0$ if the assassin does not put poison into the coffee and $P = 1$ otherwise. Second, let $A = 1$ if the bodyguard administers the antidote and $A = 0$, otherwise. Third, let $VS = 1$ if the victim survives and $S = 0$ otherwise. The structural equation for the victim's survival is $VS = \neg P \vee A$. In the actual situation we have $A = 1$, $P = 0$, and $VS = 1$, so the victim survives.

This model is structurally isomorphic to the model of the overdetermination case discussed in section 3.4. If we substitute P by $\neg T$ (the negation of the trainee's action), A by S (representing the supervising assassin's behaviour), and VS by VD, then we obtain exactly the structural equations that describe the case where both assassins simultaneously fire lethal shots at the victim.

The structural isomorphism entails—again—that in both cases the same counterfactual dependencies and independencies hold. In either case the victim's survival or death does not depend on either of the other two variables. Moreover, in either case the victim's survival or death does depend on each of the variables given that the other variable is held fixed at the non-actual value. Given that the supervisor does not shoot, the victim's death depends on the trainee's actions. The same holds for the trainee. Likewise in Bogus Prevention. Given that the assassin administers poison, the victim's survival depends on the bodyguard's actions. And given that the bodyguard does not administer antidote, the victim's survival depends on the assassins action.

In Bogus Prevention, survival depends on the bodyguard's administering antidote given that the poisoning variable is set to a non-actual value ($P = 1$). Thus, according to the HP definition, the antidote is an actual cause of the victim's survival. But this is implausible because there was never any poison that needed to be neutralized.

Thus, Bogus Prevention is a counterexample to the HP definition. Again, one way to respond to this counterexample would be to search for a refined definition that excludes the bodyguard's behaviour as an actual cause. However, the problem is that any such refinement would at the same time exclude the trainee's actions in symmetrical overdetermination as an actual cause. The reason for this is that the bodyguard and the trainee are equivalent from the perspective of the causal models.

I conclude that the Problem of Isomorphism poses a principled challenge to the kind of account that we have discussed so far. It generates counterexamples that cannot be fixed through a straightforward refinement. Any such refinement would capture the new examples at the price of loosing the old examples. The reason is that the corresponding cases are represented by isomorphic causal models. That is, at the level of causal models there is nothing to distinguish between the cases. This suggests that causal models miss something crucial about the example cases, something that would explain why they evoke conflicting causal judgements.

3.7. Defaults

In this section I will review accounts that aim to solve the Problem of Isomorphism by employing a distinction between default and deviant values of variables. Roughly speaking, a default is a value that a variable is expected to take on or that it should take on if no further information is given. The notion of default is usually associated with the notion of normality that we have discussed in Chapter 2. The idea is that a default is a value that a variable normally takes on, whereas deviant values are abnormal. Most authors do not distinguish between different kinds of normality at this stage but invoke a notion of normality that integrates descriptive and prescriptive notions.[15]

The default/deviant distinction has been implemented in a variety of ways. The first distinction can be drawn between (1) accounts that integrate the default/deviant distinction into the semantics of counterfactual conditionals and (2) accounts according to which the default/deviant distinction is a conservative extension of standard causal models.

[15] See e.g. Halpern and Hitchcock (2015). Bear and Knobe (2017) provides empirical evidence to the effect that a mixed notion of normality is operative in many contexts. However, Beckers and Vennekens (2016) propose to distinguish between statistical and prescriptive norms on the formal level.

Let us first address integrated accounts (Menzies (2004, 2007, 2009); Huber (2013)). Here the default/deviant distinction directly affects the truth conditions of counterfactuals. The idea is that actual causation (difference making, in Menzies's terminology) is to be defined directly in terms of counterfactual dependence, that is, without the restrictions discussed in the previous sections: "C *makes a difference to* E in an actual situation relative to the model M if and only if $C \ \Box\!\!\rightarrow_M E$ and $\sim C \ \Box\!\!\rightarrow_M \sim E$" (Menzies, 2004, 166). But this counterfactual is evaluated not with regard to the actual world but with regard to the most normal world.

I shall illustrate the idea by showing how this kind of account deals with the isomorphism between symmetrical overdetermination and Bogus Prevention.[16] First, consider symmetrical overdetermination. Presumably the most normal possible world is where neither the assassin in training nor the supervising assassin attack the victim and the victim survives. In this world both the actions of the trainee and the supervisor make a difference to the victim's survival and, thus, they are actual causes. Second, consider Bogus Prevention. Presumably the most normal possible world is one where the assassin does not administer any poison and the victim survives. In this world the bodyguard's actions do not make a difference to the victim's survival and, thus, the bodyguard is not an actual cause.

Integrated accounts are to be distinguished from accounts that incorporate the default/deviant distinction as a conservative extension of causal models. Conservative extension accounts (Halpern and Pearl (2005); Hitchcock (2007b); Halpern (2008); Halpern and Hitchcock (2015); Halpern (2016)) leave the semantics of counterfactual conditionals unchanged but argue that defaults and deviants determine which counterfactuals feature in our judgements of actual causation. According to Halpern and Hitchcock (2015), causal reasoners have, "in addition to a theory of causal structure (as modelled by the structural equations), a theory of 'normality'

[16] The problems arising from the isomorphism between "Backup" and "Careful Poisoning" are not as easily accounted for, see the treatment in Menzies (2017).

or 'typicality'" (433).

A key difference between the two kinds of accounts is that integrated context-sensitive theories imply one kind of causal concept while conservative extension accounts (at least sometimes) distinguish context-sensitive causal notions from causal notions that are independent of the context. Halpern and Hitchcock, for example, "envision a kind of conceptual division of labour, where the causal model $(\mathcal{S}, \mathcal{F})$ represents the objective patterns of dependence that could in principle be tested by intervening on the system, and $\succeq$ represents the various normative and contextual factors that also influence judgements of actual causation" (2015, 435, $\mathcal{F}$ denotes the set of structural equations).

How exactly does the default/deviant distinction restrict causal claims? Within conservative extension accounts (which shall be the focus in the following) there are three possible ways in which this can be achieved. First, there are accounts (e.g. Halpern and Pearl (2005)) that simply prohibit deviant variable settings. Second, there are accounts according to which the default/deviant distinction affects a normality ranking that runs over possible worlds (Halpern and Hitchcock (2013, 2015)). A third alternative is to impose such normality rankings on sets of contexts (Halpern (2016)). In this section I shall focus on normality rankings that run over possible worlds.

According to Halpern and Hitchcock (2015), a causal reasoner's theory of normality takes the form of an order over a set of worlds. A world is defined as an assignment of values to all variables in a causal model. That is, "a world is a complete description of a situation given the language determined by the set of endogenous variables" (434).

Take the three-variable model of "Backup" as an example. Table 3.1 lists all possible worlds of this model, including the actual world where the trainee shoots ($T = 1$), the supervisor does nothing ($S = 0$) and the victim dies ($VD = 1$). Note that

T	S	VD
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	**0**	**1**
1	1	0
1	1	1

Table 3.1.: Possible worlds in "Backup." The actual world is printed in bold type.

the possible worlds also include those worlds that do not conform to the structural equations. For example, there is a possible world in which both the trainee and the supervisor shoot and the victim still survives.

Normality is then represented in terms of an ordering $\succeq$ over worlds, where $s \succeq s'$ means that world s is at least as normal as world s'. The ordering is reflexive, which means that for all worlds s it is true that $s \succeq s$, that is, each world s is at least as normal as itself. Moreover, it is transitive, that is if $s \succeq s'$ and $s' \succeq s''$, then $s \succeq s''$. Finally, the order is partial in the sense that there are worlds that are incomparable. That is, there are worlds s and s' such that neither $s \succeq s'$ nor $s' \succeq s$ holds. Incomparability accounts for the fact that normality can be evaluated along multiple dimensions.

The next step is to incorporate the normality ordering into the causal model. An *extended causal model* $M = (\mathcal{S}, \mathcal{E}, \succeq)$ is a model that includes the normality ordering along with the signature and the structural equations of a standard causal model. Moreover, the definition of actual causation needs to be adjusted such that it reflects the normality ordering. Halpern and Hitchcock (2015) suggest adding the criterion that $s_{\vec{X}=\vec{x}', \vec{W}=\vec{w}', \vec{u}} \succeq s_{\vec{u}}$ to Halpern and Pearl's version of condition AC2 which yields:

AC2$_{III}$ There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that if $(M, \vec{u}) \models Z = z^*$ for all

$Z \in \vec{Z}$, then both of the following conditions hold:

(a) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}']\neg\varphi$ and $s_{\vec{X}=\vec{x}', \vec{W}=\vec{w}', \vec{u}} \succeq s_{\vec{u}}$.

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\varphi$ for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$.

For future reference it will be useful to introduce the notion of a *witness*. A witness for $\vec{X} = \vec{x}$ being a cause of φ in context $\vec{u}$ is a world s where $\vec{X}$ and $\vec{W}$ take on non-actual values such that the effect does not occur ($\neg\varphi$). With the added normality condition the witness $s_{\vec{X}=\vec{x}', \vec{W}=\vec{w}', \vec{u}}$ has to be at least as normal as the actual world $s_{\vec{u}}$.

3.8. Solving the Problem of Isomorphism

Let me put this new definition of actual cause to work and show how it is thought to solve the Problem of Isomorphism. Consider the isomorphism of "Backup" and "Careful Poisoning." We have to show that with the normality criterion the bodyguard's administering the antidote is *not* an actual cause of the victim's survival. According to Halpern and Hitchcock, one natural choice for defaults would be $A = 0$ (representing the fact that the bodyguard does not administer antidote) and $P = 0$ (representing the fact that the assassin does not administer poison). With these defaults we arrive at a normality ordering according to which the actual world ($A = 1$, $P = 1$, $VS = 1$) is less normal than the witness of the bodyguard's being an actual cause ($A = 0$, $P = 1$, $VS = 0$). That is, with this normality ordering the bodyguard's actions *are* an actual cause.

So we need a different ranking. The value of P depends on the value of A. This means, according to Halpern and Hitchcock, "that when we think about what is typical for P, we should rank not just the typicality of particular values of P, but the typicality of different ways for P to depend on A" (2015, 451). In particular, Halpern and Hitchcock suggest a ranking where the most typical situation is one where the assassin does not administer poison regardless of the bodyguard's actions ($P = 0$). This is followed by a less typical kind of situation where the assassin behaves according to the given structural equations ($P = A$). Finally, the least typical situation is one where the assassin administers poison, again, regardless of the bodyguard's actions ($P = 1$).[17]

According to Halpern and Hitchcock, this choice results in a ranking that reproduces the correct judgement. In the actual world ($A = 1$, $P = 1$, $VS = 1$) variable A takes on a deviant value and variable P takes on a value of medium typicality (because it is determined by the given structural equation $P = A$). In the witness ($A = 0$, $P = 1$, $VS = 0$) variable A takes on a default value and we have $P = 1$ regardless of A, which is the least typical value for P. Presumably it is not clear whether the gain in normality through the shift from $A = 1$ to $A = 0$ outweighs the loss of normality through the shift from $P = A$ to $P = 1$, which means that the two worlds are incomparable. Incomparability means that the normality condition in criterion AC2$_{III}$a is not fulfilled and, thus, the bodyguard's administering antidote is not classified as an actual cause of the victim's survival.

However, note that there is an alternative solution to this instance of the Problem of Isomorphism that works without the default/deviant distinction. Blanchard and Schaffer (2017) argue that an essential requirement for identifying actual causes is that the underlying causal models be appropriate. Moreover, they argue that the given instance of the Problem of Isomorphism arises because it involves a model that is not appropriate. In particular, they argue that the given model of "Careful Poisoning" leaves out essential information about whether the bodyguard's antidote neutralizes any poison or not. Including a corresponding variable yields the model displayed in figure 3.8. This causal model is no longer isomorphic to the model of

[17] What are the reasons for accepting this particular ranking? Wouldn't it be plausible to rank the abnormality of the bodyguard's violating the structural equations higher? After all the structural equations are the most reliable information that we have about the case. A more detailed discussion of this problem will be provided in Chapter 4.

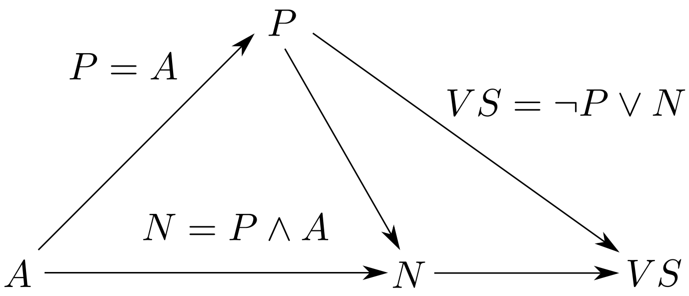

Figure 3.8.: Careful Poisoning: including a variable representing the neutralization breaks the structural isomorphism with "Backup".

"Backup." Blanchard and Schaffer (2017) use this case in order to develop a general argument against the default/deviant distinction. They argue that instances of the Problem of Isomorphism should instead be taken to indicate that one of the involved models is not an appropriate representation. This argument shall be addressed in much more detail in Chapter 7.

Next, consider Halpern and Hitchcock's (2015) account of the isomorphism between symmetrical overdetermination and Bogus Prevention. In the actual situation the assassin refrains from administering poison, the bodyguard administers the harmless antidote, and the victim survives ($P = 0, A = 1, VS = 1$). In the witness world of the bodyguard's being an actual cause the assassin administers poison, the bodyguard does not administer antidote and the victim dies ($P = 1, A = 0, VS = 0$). Halpern and Hitchcock assume, again, that refraining from action is more typical than action for both the assassin and the bodyguard. That is, $P = 0$ and $A = 0$ are the defaults. With these defaults the actual world and the witness are incomparable and, thus, the bodyguard's actions do not qualify as actual cause.

But, again, note that this problem is solved without the default/deviant distinction if we use a model that includes a variable representing the neutralization process (figure 3.9, see Halpern and Hitchcock (2015, 444) and Blanchard and Schaffer (2017, 201)). This model is not isomorphic to symmetrical overdetermination. Moreover, the bodyguard's actions are not an actual cause of the victim's survival according

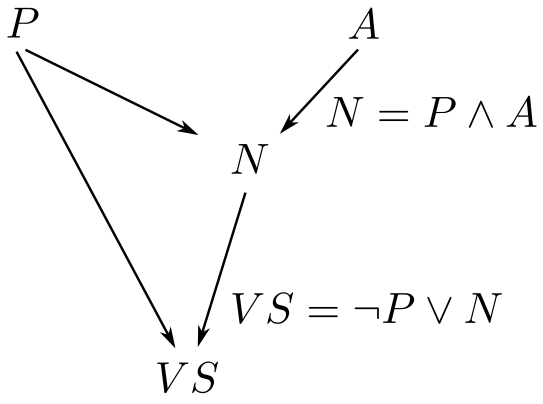

Figure 3.9.: Bogus prevention: including a variable representing the neutralization breaks the structural isomorphism with the overdetermination case.

to the definitions discussed earlier.

Blanchard and Schaffer (2017) also argue that incorporating the default/deviant distinction is problematic for independent reasons. In particular, the specific choices of what counts as default behaviour often seem to be *ad hoc*. For example, with regard to Halpern and Hitchcock's treatment of the poisoning cases there may well be reasons to believe that there are senses in which the assassin's administering poison *is* typical. After all this is what assassins normally do and what they are being paid for.

3.9. Defaults and the Problem of Selection

In the foregoing sections we have seen that the default/deviant distinction has been motivated by the Problem of Isomorphism. However, the distinction has also been considered as formalizing a solution to the *problem of selection*, introduced in Chapter 2. Let me show how Halpern and Hitchcock (2015) employ the normality ordering over possible worlds in order to account for the fact that we typically identify a lit match as an actual cause of a fire but not the presence of oxygen.[18]

[18] Similar considerations apply to the situation described in Knobe and Fraser's (2008) pen vignette, which I have discussed in section 2.7.

The situation can be represented with a causal model consisting of three variables as follows: first, there is a variable, M, which takes value 1 if the match is lit and value 0 if it is not lit. Second, there is a variable, O, which takes value 1 if oxygen is present and value 0 if no oxygen is present. Finally, there is a variable, F, which takes value 1 if a fire occurs and 0 if not. The fire depends conjunctively on the match and the presence of oxygen, that is, we have the structural equation $F = M \wedge O$.

According to the HP definition, both the lit match and the presence of oxygen qualify as actual causes of the fire. Thus, the definition does not reflect the intuitive difference between the two factors. However, taking into account the normality ordering we will see the difference. According to Halpern and Hitchcock, the following is a plausible ordering:

$$(M = 0, O = 1, F = 0) \succeq (M = 1, O = 1, F = 1) \succ (M = 1, O = 0, F = 0)$$

According to this ordering, the witness world of the match being the actual cause (that is, the world where oxygen is present but no match is lit such that there is no fire) is at least as normal as the actual world. But the witness world of the presence of oxygen (that is, the world where the match is lit but no fire occurs because of the absence of oxygen) is less normal than the actual world. The reason for this is that the presence of oxygen is the default state.

Halpern and Hitchcock also emphasize the contextual character of the normality ordering. If the fire occurs in a place that is typically voided of oxygen (e.g. in a laboratory), then the normality ordering would be different such that the witness of oxygen being an actual cause would be at least as normal as the actual world. Thus, the account nicely reproduces the contextual character of selective causal judgements discussed in Chapter 2.

3.10. Conclusion

In this chapter I have reviewed causal model approaches to actual causation. The Halpern-Pearl definition seems to be a particularly promising candidate because it accounts for cases involving early and late preemption as well as for cases involving symmetrical overdetermination, which all have been difficult to handle for counterfactual theories of causation. Moreover, adding Halpern and Hitchcock's (2015) normality condition as a restriction to the Halpern-Pearl definition seems to yield an approach that accounts also for the problem of selection. Thus, it seems like causal models provide the framework for a unified solution to the problems of redundancy and selection.

Part II.

Pluralism about Actual Causation

4. Pluralism about Actual Causation

A key assumption underlying the debate that I have reviewed in the foregoing chapter is that there is a unified concept of actual causation.[1] As an example, consider Halpern's *Actual Causality* (2016). Surveying a range of toy examples Halpern offers three competing basic definitions of actual causation and discusses several ways of implementing considerations of normality. At the same time, Halpern seems to be convinced that there is one unified concept of actual causation that captures the causal intuitions in the wealth of example cases and he suggests that his "modified" concept of actual causation comes closest. Moreover, even authors who endorse causal pluralism seem to subscribe to the view that the concept of actual causation is a unified concept among the plurality of other causal concepts. For example, in *Making Things Happen* (2003), chapter 2, Woodward uses causal models in order to clarify the relation between causation and manipulability. The result is a theory that distinguishes between total causes, direct causes, contributing causes, and actual causes. Thus, while Woodward accepts that there is a plurality of causal concepts, he seems to assume that within that plurality 'actual causation' refers to a unified concept.

In this chapter I will challenge the assumption that there is a unified concept of actual causation and suggest a pluralist theory instead. In particular, I propose that we distinguish between total, path-changing, and contributing actual causes.

[1] Chapter 4 builds on my article "Three Concepts of Actual Causation" (Fischer (forthcoming a)) accepted for publication by *The British Journal for the Philosophy of Science* on 03/30/2021 (`https://doi.org/10.1086/715201`).

Total actual causation amounts to a straightforward counterfactual dependence of the effect on the cause. Path-changing actual causation amounts to counterfactual dependence of the effect on the cause given that certain other variables are held fixed at their *actual* values. Thus, path-changing actual causation is similar to the definitions of actual causation provided by Pearl (2000), Hitchcock (2001), and Halpern (2015). Finally, contributing actual causation amounts to counterfactual dependence of the effect on the cause given that certain other variables are set *non-actual* values and, thus, is similar to the concept captured by Halpern and Pearl's (2005) definition of actual causation.

I will provide two lines of argument for my pluralist account. The first line of argument is based on three example cases that shall illustrate the problems that unified accounts of actual causation face. The examples have in common that there are two factors that both seem to qualify as actual causes. Yet, at the same time, there is an important asymmetry between these factors. Extant unified theories face a dilemma here: either they describe both factors as actual causes (and thus cannot explain the asymmetry), or they dismiss the intuition that both factors are actual causes (and account for the asymmetry by identifying only one of the factors as actual cause). By distinguishing between total, path-changing, and contributing actual causes, my account can both hold that each example involves two actual causes, and explain the perceived asymmetry. For example, in trumping cases like the case where the major and the sergeant order the corporal to advance (see Chapter 2) both the major's and the sergeant's order appear to be actual causes because the respective causal processes both run to completion. Nevertheless, the major differs from the sergeant in that the major has more control over the situation (he could have ordered the corporal to retreat). Extant accounts of actual causation do not reflect this intuition in an appropriate way.[2] On the one hand, there are

[2] There is an exception: Hitchcock's (2011) contrastive account of actual causation in trumping cases provides a solution that is very similar to the one that will be provided here. Even though Hitchcock

accounts that identify only the major as actual cause and, thus, account for the perceived asymmetry between the two officers (such as Lewis's influence account). On the other hand, there are accounts that identify both the major and the sergeant as actual causes but then treat them on a par (such as Halpern and Pearl's (2005) account). My pluralist theory accounts for the relevant intuitions by identifying the major and the sergeant both as contributing actual causes but only the major as a total actual cause.

The second line of argument is based on a functional approach to actual causation. Following Woodward (2014), a functional approach, among other things, informs us about relevant distinctions among causal concepts by considering the purpose of such concepts. In particular, I will argue that if we take concepts of actual causation to inform intervening agents about factors that are particularly suited for intervention, then the distinctions between total, path-changing, and contributing actual causation are important. The distinction between total and path-changing actual causation, for example, is important from the interventionist perspective because it informs the agent about how control about an effect can be achieved. Total actual causation can be exploited by simply targeting the total actual cause. Path-changing actual causation, by contrast, requires that the intervention on the path-changing actual cause be combined with a secondary intervention that counteracts the adverse consequences of the first intervention.

After presenting these two lines of argument I will examine the role of normality in my pluralist account of actual causation. In particular, I will argue that we need to discern two kinds of context-sensitivity. First, there are considerations that concern the values of individual variables (context-sensitivity$_1$). Second, there are considerations that concern violations of structural equations (context-sensitivity$_2$).

The chapter is structured as follows. In section 4.1 I will introduce three example

does not explicitly endorse a pluralism about actual causation in the article, his discussion involves distinctions between concepts of actual causation that are similar to the ones discussed here.

cases that raise problems for unified accounts of actual causation: trumping cases, the Light Bulb case, and the Henchman case. In particular, I will show that the HP definition fails to explain an intuitive asymmetry that is involved in each of these cases. Moreover I will argue that, unlike the asymmetries discussed in the foregoing chapter, these asymmetries are *not* explained by considerations of normality. In section 4.2 I will provide definitions of total, path-changing, and contributing actual causation. I will show how the distinction between these concepts helps account for the problem cases described in section 4.1 and I will relate the definitions to differences between concepts of causation that have been drawn in the literature. In section 4.3 I will turn to the second line of argument and provide a functional justification for the distinction between total, path-changing, and contributing actual causes. Finally, in section 4.4 I will address the role of norms within this pluralist account of actual causation and introduce the distinction between two kinds of context-sensitivity.

4.1. Three Problem Cases

In this section we shall have a look at three example cases that pose problems for the unified accounts discussed in the foregoing chapter. In the following discussion I will focus on Halpern and Pearl's (2005) account because I take this account to provide the most advanced treatment of redundancy cases. I take the argument to extend to the other accounts presented in the foregoing chapter and I will indicate relevant differences as we go along.

4.1.1. Trumping

A major and a sergeant order a corporal to advance. The corporal receives both orders at the same time and advances. What caused the corporal to advance? The case is ambiguous. On the one hand, the corporal receives both orders at the same time. Thus, it seems like both orders are causes just as in situations of symmetrical overdetermination. On the other hand, orders of higher-ranked officers trump those of lower-ranked officers. Thus, it seems like there is an important asymmetry between the two orders such that some have argued that the only cause is the major's order (Schaffer (2000a); Lewis (2004)).

Halpern and Pearl (2005) are aware of the ambiguity of trumping cases and they offer two kinds of models. The first model is a coarse-grained model that consists of three variables (see figure 4.1A for the DAG): there are two variables, M and S, representing the major's and the sergeant's actions, respectively. These have three possible values: 1 (order advance); −1 (order retreat); 0 (do nothing). Moreover, there is a variable, A, representing what the corporal does. If the major issues an order, then the corporal follows it, that is, $A = M$ if $M \neq 0$. If the major does not issue an order then the corporal follows the sergeant's orders, that is, $A = S$ if $M = 0$. In the actual situation both the major and the sergeant order to advance and the corporal advances: $M = S = A = 1$. With this model the HP definition identifies both the major and the sergeant as actual causes of the corporal's action.

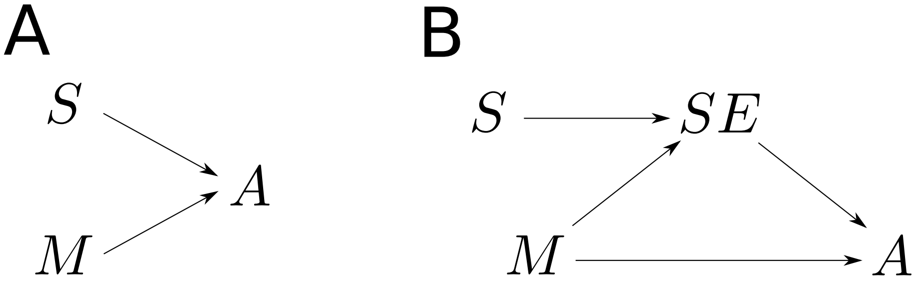

Figure 4.1.: Two models for trumping preemption.

But what about the perceived asymmetry between the major and the sergeant? Halpern and Pearl (2005) argue that there is an alternative model according to which only the major is an actual cause. The alternative model includes a variable SE that captures whether the sergeant's order is effective. If the major does nothing ($M = 0$), then $SE = S$. But if the major issues an order ($M \neq 0$), then the sergeant's order is not effective, a fact represented by $SE = 0$ (see figure 4.1B for the DAG).

According to Halpern and Pearl, this model reproduces Schaffer's and Lewis's claim that only the major's issuing the order is an actual cause of the corporal's actions. However, this does not seem to be a legitimate treatment of trumping cases, at least if they are construed as cases where both causal processes run to completion. In fact, Schaffer anticipates this kind of treatment of trumping cases. In reply he argues that the judgement that the major is an actual cause but not the sergeant seems to arise independently of considerations regarding such sophisticated assumptions about the corporal's decision module. Moreover, Halpern and Pearl's sophisticated model does not apply to cases like the Merlin and Morgana case which is set up in a way that makes a treatment along these lines impossible.

But even if this treatment were legitimate, Halpern and Pearl's account would not be completely satisfactory. According to the coarse-grained model, both officers are actual causes and, according to the fine-grained model, only the major is an actual cause. That is, according to these models, it is *either* the case that the officers are to be treated on a par *or* it is the case that only one of them is an actual cause. But the intuition seems to be that both officers are actual causes *and* that there is an asymmetry between them.

Could the perceived asymmetry be related to considerations of normality? First of all, the case seems to be relatively neutral as regards considerations of normality. The case does not provide any specific information about whether any of the involved officers act according to any relevant norms or whether they violate any relevant norms. But maybe norms are at play after all. Then their influence should become clearer if the case is more drastic. So, suppose the orders require the corporal to violate a moral norms. Then the major is an actual cause because there is a witness world which is more normal than the actual world ($(M = 0, S = 0, C = 0)$, that is, the world where none of the officers issue an order). But the same world is also a witness for the sergeant's being an actual cause. Thus, it seems that the perceived asymmetry is not to be explained by considerations of normality.

4.1.2. The Light Bulb

Here is another problematic case:

> "A and B each control a switch. There are wires going from an electricity source to these switches and then continuing on to C. A must first decide whether to flip his switch left or right, then B must decide (knowing A's choice). The current flows, resulting in a bulb at C turning on, iff both switches are in the same position. B wants to turn on the bulb, so flips her switch to the same position as A does, and the bulb turns on" (Halpern 2016, 100).

Figure 4.2A depicts a schematic representation of the kind of circuit that is described here. The structural equations of the case, following Halpern, are: $B = A$, saying that B copies the position of A's switch, and $C = 1$ iff $A = B$, saying that the light bulb is switched on if and only if the two switches are in the same position.

Causal intuitions are again ambiguous. On the one hand, both agents' actions are causes of the fact that this particular circuit is closed. On the other hand, there is clearly an asymmetry between the two agents because only B has control over whether the light is switched on. This ambiguity is also reflected by Halpern's discussion. On the one hand, Halpern states that "[i]ntuition suggests that A's action should not be viewed as a cause of the C bulb being on, whereas B's should" (ibid.). On the other hand, his treatment of the case in terms of normality (see the discussion in section 4.1.4) suggests that he takes both agents to be causes, while the perceived asymmetry between them is a matter of degree and is to be explained

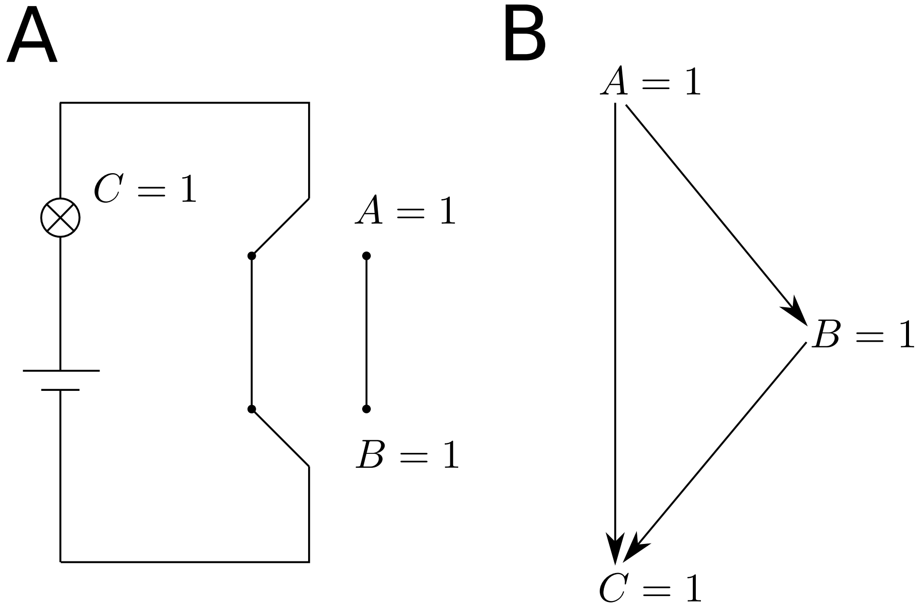

Figure 4.2.: The Light Bulb. If both switches are either on the left or right side the lamp C is switched on. A: the circuit. B: the causal graph.

in terms of considerations associated with normality. The HP definition (as well as Halpern's modified definition that is the focus of his discussion of the example) treats both agents on a par and, thus, does not reflect the important fact that only *B* has control over the lamp's being switched on.

Is the asymmetry explained by considerations of normality? The case appears to be relatively neutral with regard to normality. Thus, we should not expect that the intuitive asymmetry is related to considerations of normality. Nevertheless, a good way to test whether norms are at play is to make the case more drastic. We shall replace the event of a lamp's being switched on by the morally abnormal event of a person getting an electric shock (as in the structurally equivalent scenario known as 'Shock C' (McDermott (1995))). But even in this case the difference between *A* and *B* does not seem to be explained by normality, if implemented in the way discussed in Chapter 3. The world where $A = 0$, $B = 1$, and $C = 0$ is a witness for *A* being an actual cause and it seems to be more normal than the actual world. But the same

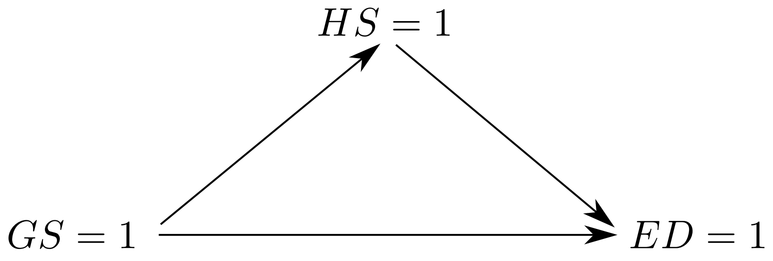

Figure 4.3.: The Henchman. The henchman copies the gangleader's actions. The victim dies if either the gangleader or the henchman shoots.

world is also a witness for B. Thus, both A and B qualify as actual causes.

4.1.3. The Henchman

Consider the following case, taken from Rosenberg and Glymour's (2018) review of Halpern's *Actual Causality*. A gang leader ($GS = 1$) and his henchman ($HS = 1$) both shoot and their enemy dies ($ED = 1$). The enemy would also have died if either only the gang leader or only the henchman had shot. However, the henchman shoots if and only if the gang leader shoots. We shall also assume that the two individually lethal bullets kill instantaneously and hit at the same time, such that they stand in a relation of symmetrical overdetermination. The causal graph is displayed in figure 4.3 and the structural equations are as follows:

$$HS = GS$$

$$ED = GS \vee HS$$

Clearly, both the gang leader and the henchman are to be identified as actual causes of the enemy's death. Yet, at the same time there seems to be an important asymmetry between the two agents because the enemy's death depends upon the gang leader's actions but not on the henchman's actions.

Rosenberg and Glymour present this case as a counterexample to Halpern's modified definition. According to Halpern's modified definition, the gang leader is an actual cause (because of the relation of counterfactual dependence) but the henchman is not an actual cause because there is no way to restore the counterfactual dependence of the enemy's death on the henchman's action by keeping the values of other variables fixed at their actual values. Moreover, the henchman is not 'part of a cause' according to Halpern's modified definition, because $\{HS = 1, GS = 1\}$ violates the minimality condition: $GS = 1$ by itself is an actual cause.

The HP definition does identify both the gang leader and the henchman as actual causes. However, it treats both agents on a par and, thus, does not explain the intuitive difference between the two agents. Neither does consideration of normality explain the difference between the two agents. The only witness for the gang leader being an actual cause is the world where none of the agents shoot ($GS = 0, HS = 0, ED = 0$). But this is also the only witness world for the henchman. Thus, if one of the two agents qualifies as actual cause, according to the normality criterion, then the same should hold for the other agent.

4.1.4. Graded Causation to the Rescue?

The problem with the HP definition in the three examples is that it does not reflect the intuitive asymmetry between the involved agents. Moreover, we have seen that applying Halpern and Hitchcock's (2015) normality criterion does not help distinguish between the agents. However, there have also been suggestions to use normality in order to implement a graded notion of actual causation. *Prima facie* this could lead to a plausible solution to the problems because the difference between the agents does not seem to be a difference between causes and non-causes but a difference in degree.

One way to implement a graded notion of actual causation is to rank causes according to the normality of their most normal witness (Halpern and Hitchcock, 2015, 436). According to this approach, both the short-circuit and the presence of oxygen are actual causes of the fire. But the short-circuit is a *better* cause because its most normal witness (a world like the actual world but where the short-circuit does not occur) is more normal than the most normal witness of the presence of oxygen (a world like the actual world but without the presence of oxygen).

First, consider the variation of the trumping case where the officers issue norm-violating orders. Presumably, the most normal world then is the one where none of the orders is issued and the corporal does not do anything. But this is the best witness for both the major and the sergeant, meaning that graded causation does not help distinguish the two. Second, with regard to the Light Bulb case the witness of A is ($A = 0$, $B = 1$) and the witness of B is ($A = 1$, $B = 0$). Both worlds seem to be less normal than the actual world because both worlds involve a violation of a structural equation. But apart from this there does not seem to be a difference in the normality of these two witnesses that could explain the perceived asymmetry between A and B. Finally, in the henchman case there is only one witness world for the gang leader being an actual cause ($GS = 0, HS = 0, ED = 0$). But this is also the only witness for the henchman being an actual cause. So, again, graded causation does not explain the difference between the agents.

Halpern (2016) proposes yet another way to implement (graded) normality. With regard to the Light Bulb case Halpern argues that

> we can take the change from the world where both $A = 1$ and $B = 1$ to the world where $A = 1$ and $B = 0$ to be smaller than the one to the world where $A = 0$ and $B = 1$, because the latter change involves changing what A does as well as violating normality (in the sense that B does not act according to the equations), while the former change requires only that B violate normality. This gives us a reason to prefer[3] $B = 1$ as a cause (2016, 102).

[3] A mere preference for B does not exclude that A is a cause as well. Thus, Halpern seems to assume (rightly, I think) that both factors are causes but to a different degree.

In order to show that $A = 1$ is an actual cause (according to Halpern's modified definition as well as the HP definition) we need to apply two interventions. First, we need to intervene on A and, second, we need to intervene on B such that it keeps its actual value. In order to show that $B = 1$ is an actual cause, however, we have to apply only one intervention on B. Halpern takes this single intervention to amount to a smaller decrease in normality than the combination of interventions that is required to show that A is an actual cause.

In the following we will see that the distinction between applying one rather than two interventions is indeed essential. Thus, I think that Halpern's discussion points to an important difference between A and B.[4] However, Halpern employs the number of interventions in order to implement a *difference in normality* between A and B. And this seems to be problematic. The idea of normality was to rank different interventions. The short-circuit is a better actual cause of the fire because preventing the short-circuit leads to a world that is more normal than preventing the presence of oxygen. But with Halpern's new criterion it seems that the short-circuit and oxygen are to be treated on a par because arriving at either factor's witness requires one intervention (see Rosenberg and Glymour (2018) for a similar criticism of Halpern's normality criterion). I conclude that none of the attempts to implement a graded notion of actual causation via a criterion of normality helps with regard to the three problematic examples.

[4] Similar considerations apply to the trumping case and the Henchman case, as will be argued in the following section.

4.2. Total, Path-changing, and Contributing Actual Causation

4.2.1. Definitions

The ambiguity of the three example cases suggests that there are different concepts of actual causation at play. Here are definitions of three concepts of actual causation.

Total Actual Cause

$X = x$ is a total actual cause of φ in $(M, \vec{u})$ iff

TAC1: $(M, \vec{u}) \models (X = x) \wedge \varphi,$

TAC2: There exists a setting x' of variable X such that

$$(M, \vec{u}) \models [X \leftarrow x']\neg\varphi. \tag{4.1}$$

The definition of total actual cause amounts to straightforward counterfactual dependence of the effect on the cause. Condition TAC1 says that in the actual situation both the cause and the effect have to be instantiated and condition TAC2 expresses the requirement of counterfactual dependence. The concept of total actual causation describes an important subclass of factors that are typically identified as actual causes, yet, it surely does not capture all such factors since it does not apply to cases involving redundancy.

Path-changing Actual Cause

$X = x$ is a path-changing actual cause of φ in $(M, \vec{u})$ iff

PAC1: $(M, \vec{u}) \models (X = x) \wedge \varphi.$

PAC2: There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting x' of variable X such that if $(M, \vec{u}) \models \vec{W} = \vec{w}^*$, then

$$(M, \vec{u}) \models [X \leftarrow x', \vec{W} \leftarrow \vec{w}^*]\neg\varphi.$$

This definition formalizes the idea that effects depend on their path-changing actual causes given that certain other variables are being kept fixed at their actual values. The definition captures the idea that preempting factors, such as the assassin in training in Backup, are actual causes. But we will see that the definition also helps account for cases where the path-changing actual cause does not preempt the competing factors, as in the light bulb case. I choose the label 'path-changing actual cause' because intervening on such a cause changes the causal path along which the effect is influenced. If I intervene on switch A's switch, then I will not prevent the light's being switched on. But as a result of the intervention the electric current will run along a different wire. Note that the definition of path-changing actual cause is wider than the definition of total actual cause, that is, every total actual cause is a path-changing actual cause (with $W = \emptyset$), but not vice versa.

The definition is inspired by and closely related to Pearl's (2000), Hitchcock's (2001), and Halpern's (2016) definitions of actual causation. But let me point out some differences. The definition differs from Hitchcock's active route criterion because it does not require that there is a *single* active route. Thus, it avoids problems associated with cases where the influence is transmitted via several paths that are active at the same time as in the example involving the chief assassin discussed in section 3.3. Second, the definition is different from Halpern's definition of modified actual causation because it accepts only single variables $X = x$ as actual causes but not sets of variables $\vec{X} = \vec{x}$. The restriction is motivated by the fact that typically we take the single variables of a causal model to be independently manipulable factors.

Treating a set of variables as actual causes then seems to neglect the important fact that several interventions are to be applied in order to avoid the outcome.[5] One consequence of restricting the definition as applying to individual variables is that it does not account for cases of symmetrical overdetermiantion—this seems to be Halpern's motivation for endorsing the view that sets of variables can count as an actual cause. But this is not a problem because for this purpose we have the concept of contributing actual cause.

Contributing Actual Cause

$X = x$ is a contributing actual cause of φ in $(M, \vec{u})$ iff

CAC1: $(M, \vec{u}) \models (X = x) \wedge \varphi$

CAC2: There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $X \subseteq \vec{Z}$ and some setting $(x', \vec{w}')$ of the variables in $(X, \vec{W})$ such that if $(M, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then both of the following conditions hold:

(a) $(M, \vec{u}) \models [X \leftarrow x', \vec{W} \leftarrow \vec{w}']\neg\varphi$.

(b) $(M, \vec{u}) \models [X \leftarrow x, \vec{W}' \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\varphi$ for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$.

Note that the concept of contributing actual cause is strictly more encompassing than the concept of path-changing actual cause. Every path-changing actual cause is also a contributing actual cause, but not vice versa. More specifically, path-changing causaution describes the special case where the values of variables in $\vec{W}$ may only be set to values $\vec{w}^*$ that these variables take on in the actual situation.

The definition of contributing actual cause is almost identical to Halpern and Pearl's (2005) definition of actual cause. The only difference is that, again, my definition is restricted to individual variables as actual causes.

[5] Schaffer (2003) gives a systematic defense of this view, which he calls individualism.

4.2.2. Trumping, Light Bulb, and Henchman Reconsidered

Let us see how the distinction between these concepts helps account for the problem cases of the foregoing section. We shall begin with trumping cases. Consider Halpern and Pearl's simple model of the case. According to this model, both the sergeant and the major are identified as contributing actual causes. Given that the major does not issue an order, the corporal's behaviour depends upon the sergeant's order. The same is true for the major. Given that the sergeant does not issue an order, the corporal's actions depend on the major's actions. However, there are also actions available to the major where the dependence holds regardless of the sergeant's actions—in fact, the only case where the dependence is broken is in the case where the major decides to issue no order at all. More specifically, the witness $(M = 0, S = 0, A = 0)$ indicates that the major is a contributing actual cause. But there is also a witness $(M = -1, S = 0, A = -1)$, according to which the major is a total actual cause. Thus, we can accommodate the intuition that both officers are actual causes and the perceived asymmetry is explained by the fact that only the major is a total actual cause.[6]

Halpern and Pearl appear to endorse a similar account of the asymmetry. They point out that the major is a strong cause while the sergeant isn't (2005, p. 874). The definition of strong cause builds upon the definition of contributing actual cause (or actual cause in Halpern and Pearl's terminology) and involves an extra condition requiring that $(M, \vec{u}) \models [X \leftarrow x, \vec{W} \leftarrow \vec{w}''](Y = y)$ for all settings $\vec{w}''$ of $\vec{W}$. The major's ordering to advance $(M = 1)$ is a strong cause because for all possible

[6] A similar account of trumping cases has been provided by Hitchcock (2011) who employs a contrastive notion of causation. According to Hitchcock, the major's ordering the corporal to advance *rather than ordering the corporal to retreat* is a non-redundant cause of the corporal's advancing *rather than retreating*. By contrast, the major's ordering the corporal to advance *rather than issuing no order at all* is a redundant cause of the corporal's advancing *rather than doing nothing*. My account is contrastive as well: the fact that $M = 1$ is a contributing actual cause depends on the fact that 0 is an alternative value of variable M and the fact that $M = 1$ is a total actual cause depends on the fact that -1 is an alternative possible value.

values of the sergeant variable S the corporal advances if $M = 1$. By contrast, the sergeant is not a strong cause because the corporal does not advance if the major orders to retreat.

However, one difference between my account and Halpern and Pearl's account is that they reject the concept of strong cause where it does not coincide with their concept of actual cause: 'in many of our examples, causality and strong causality coincide. In the cases where they do not coincide, our intuitions suggest that strong causality is too strong a notion' (2005, p. 855). Just as Halpern and Pearl's concepts of strong cause and actual cause my concepts of TAC and CAC coincide in most situations —situations that involve straightforward counterfactual dependence of the effect on the cause. Moreover, in instances where CAC and TAC do not coincide, the concept of TAC seems to be too restrictive to capture all those factors that are intuitively identified as causes. But I do not take this to be a reason to endorse the concept of CAC instead of the concept of TAC. Instead I argue that we need both the concept of TAC and the concept of CAC. Endorsing both these concepts is what enables us to accommodate the intuitions that apply to cases like trumping, where two factors are causes in the weaker sense of CAC and only one factor is a cause in the stronger sense of TAC.

There is a different question of whether the distinction between actual cause and strong cause can give rise to a viable pluralist account if one endorses both concepts, even in cases where the two concepts do not coincide. I think that such a pluralist account could in principle be developed. However, I prefer a pluralist account in terms of TAC, PAC, and CAC because it also accounts for the henchman case and the light-bulb case, whereas the distinction between actual and strong causation does not help in those examples.

I have argued that my account explains a perceived asymmetry between the major and the sergeant. But presumably this will not satisfy authors like Lewis (2004) and

Schaffer (2004) who claim that the sergeant is not a cause at all. Note that Lewis and Schaffer do not just have a different way of accounting for certain basic causal intuitions. They also have a different view on what the basic causal intuitions are in the first place. But how do we know which are the 'correct' causal intuitions? In section 4.3 I will address these concerns, arguing that identifying the sergeant as (contributing) actual cause makes sense from the perspective of an agent who aims to influence whether the corporal advances or not.

Next, turn to the Light Bulb case. B is a total actual cause of the light's being switched on because there is a relation of straightforward counterfactual dependence. A is also an actual cause, however, it is only a path-changing actual cause: we need to keep B fixed at its actual value, in order to reveal a counterfactual dependence of C on A. In order to see the path-changing character of A in the causal model we need to choose a model that reflects whether the current runs along the left wire (LW) in figure 4.2A or along the right wire (RW), for example, a model with structural equations as follows: $B = A$, $LW = A \wedge B$, $RW = \neg A \wedge \neg B$, and $C = LW \vee RW$. With this we can account for the intuition that both A and B are actual causes for the light's being switched on but at the same time we can account for the intuition that B has better control over the lamp's being switched on.

Finally, consider the Henchman case. The gang leader is a total actual cause of the enemy's death. The henchman is also an actual cause but he is only a contributing actual cause: we need to set GS to a non-actual value in order to reveal the dependence of the enemy's death on the henchman's actions.

4.2.3. Relation to Existing Accounts

The definitions of total, path-changing, and contributing actual cause are not novel (apart from slight modifications as noted). What is new about my account, however, is that I do not consider these definitions to be competing with each other. Instead, I take them to specify different and equally justified concepts of actual causation.[7] In order to further clarify my account I shall now point out the differences to existing pluralist accounts.

My approach has similarities with a number of pluralist proposals. Hall (2004), for example, argues that we need to distinguish between causation as dependence and causation as production. The dependence notion, according to Hall, accounts for the intuition that effects depend on their causes and can be applied to events as well as omissions. The dependence notion maps straightforwardly onto my concept of total actual causation. The production notion, according to Hall, accounts for the intuition that causation is transitive, local, and intrinsic. Hall argues that the production notion is in principle incompatible with the intuition that effects depend on their causes, especially in situations involving omissions. Even though the production notion seems to be the salient notion in examples involving redundancy such as those discussed above, I do not see a straightforward way to map it onto either one of my concepts of path-changing actual cause or contributing actual cause.

It seems like Hall could accommodate the problem cases discussed in section 4.1, arguing that in each of the cases both factors are producers but only one of them involves dependence. But it is easy to change the examples such that they generate problems for Hall. One could, for example, make the henchman case a case involving omissions. Suppose a child is drowning in a lake and any attempt to save the child would be extremely risky because there is a violent thunderstorm. There is a chief lifeguard and an assistant lifeguard who can save the child only if they go out on the water together. The chief lifeguard is more experienced and

[7] Interestingly, Pearl (2000) seems to suggest a similar distinction. He defines a notion of actual cause (similar to my concept of path-changing actual cause) and a notion of contributing cause (similar to my concept of contributing actual cause). But he does not provide an explicit defence of a pluralism about actual causation and in more recent publications (e.g. Halpern and Pearl (2005)) this distinction is not maintained. Instead, the HP definition is taken to capture the notion of actual cause.

her assistant always follows her decision. Because of the thunderstorm the chief lifeguard decides to stay on land, and so does the assistant. As a result the child drowns.

My theory says that both lifeguards are actual causes. The chief lifeguard is a total actual cause and the assistant is a contributing actual cause, analogous to the henchman case. According to Hall, the chief lifeguard is a cause in the sense of the dependence notion, but the assistant lifeguard is not a cause at all—which strikes me as unintuitive. The problem is that the lifeguards cannot be described as producers because the child's drowning results from omissions rather than actions.

A further difference between Hall's and my approach is that my account is more optimistic with regard to causal models as a framework for defining concepts of actual causation. Hall thinks that a counterfactual analysis can be given only of his dependence notion of causation, while for the production concept we need a different kind of basic building block. Applying a terminology introduced by Hitchcock (2007c) one could say that Hall proposes an extramural pluralism, meaning that he employs distinct kinds of basic building blocks (counterfactual conditionals, regularities) in order to generate a plurality of causal concepts. Then my account would be closer to an intramural pluralism because it generates a plurality of causal concepts employing only one sort of basic building block: (sophisticated) counterfactual conditionals framed in terms of causal models.

It will also be useful to compare my account with Cartwright's (2007) pluralist account. Cartwright argues that there is a wide variety of causal relations that is reflected by 'content-rich causal verbs', such as 'compress', 'attract', and 'discourage'. Cartwright states that on a unificationist view all these causal verbs are replaced by the abstract terms 'cause' and 'prevent' or, even worse, 'by one single piece of notation—the arrow [of a causal graph]' (2007, p. 21). This, she argues, is a problem because the content of the rich causal verbs is lost on such a view.

I agree with Cartwright that reducing all causal concepts to the kind of structural dependence that is expressed by the arrows of a causal graph is problematic. But I do not think that this is a problem for causal models. In fact, I take my account to illustrate how the framework of causal models can be employed to define a plurality of causal concepts. These causal concepts are surely not as specific as Cartwright's content-rich causal verbs. But neither are they as abstract as the arrows in a causal graph. Instead, they are located at an intermediate level of abstraction, a level that I take to be abstract enough to be applicable to a wide range of circumstances and, at the same time, specific enough to give an agent clear guidance with regard to suitable targets of intervention (see the argument in section 4.3).

There is another aspect of Cartwright's pluralist account, which is similar to the extramural pluralism of Hall. Cartwright argues that there are multiple theoretical frameworks that each account for certain paradigmatic features of causal reasoning but that each framework also has its limitations. She argues that the structural model framework, for example, faces problems where its central assumption of modularity does not apply (Cartwright, 2007, p. 13). Her example is a carburettor. She argues that the entanglement of several causal relations in this kind of system is even an explicit aim of efficient engineering.[8]

There are two ways the intramural pluralist can respond to this kind of worry. The intramural pluralist can attempt to provide an account that exhausts all kinds of causal concepts such that extramural pluralism would not be needed. Alternatively, the internal pluralist can develop an intramural theory that tries to get as far as possible, but also acknowledge that there may remain instances of causation that

[8] However, consider the criticism by Steel (2010): Cartwright's carburettor example only shows that there are *some* interventions that are non-modular, namely those on the carburettor's geometry. This shouldn't pose a problem for the interventionist account because with respect to almost any system it is easy to find some non-modular intervention. What Cartwright needed to show is that there are no modular interventions. And this does not seem to be true since we can intervene independently on other parts of the carburettor such as the air filter, the choke value, and the throttle valve.

will need other theoretical frameworks. On this view intramural and extramural pluralism coexist.

I take intramural pluralism to be able to make genuine progress even if it coexists with extramural pluralism. In analogy to David Chalmer's (1996) hard and easy problems of consciousness, Hitchcock (2007a) distinguishes between hard and easy problems of causation. The hard problems of causation concern among other things "the origins of the distinction between the causal and the non-causal" (2007a, 58). Hitchcock argues that there may be philosophical progress on these problems of causation but that we "should not expect these hard problems to go away any time soon" (ibid.). The easy problems concern "distinctions among the different kinds of causal relationships" (ibid.). Hitchcock argues that investigating into these distinctions is a place where we can hope to make genuine philosophical progress. Like Chalmer's easy problems of consciousness the easy problems of causation are still hard, but "they should at least be tractable" and he concludes that "[t]he real work that [...] needs to be done is that of providing useful taxonomies for causal relationships" (59). I consider my project to be a contribution to the project of providing such useful taxonomies.

An advantage of my intramural account (over a purely extramural account) is that it clarifies the relationship between the different concepts of actual causation needed to account for cases involving redundancy. This is possible because the concepts are all defined in the same formal framework, that ultimately relates them to underlying relations of counterfactual dependence. The concept of CAC is the most permissive concept while the concepts of PAC and TAC describe special cases. More specifically, the three concepts exhibit a nested structure: TAC is a special case of PAC, and PAC is a special case CAC. This implies that every TAC is also a PAC and CAC, and that every PAC is a CAC—just as every square is a rectangle, and every rectangle is a quadrilateral (a plain figure with four edges and four vertices).

But doesn't the hierarchical relation between CAC, PAC, and TAC threaten the status of PAC and TAC as independent concepts of actual causation? That is, couldn't we state that the concept of CAC provides the basic analysis of what it means for an event to be an actual cause, and doesn't this imply that we have a unified concept of actual causation after all? It is true that the concept of CAC is the most general concept. However, this does not make the concepts of PAC and CAC obsolete. I take the moral of the foregoing section to be that the concept of CAC (as well as the concept of PAC), if taken in isolation, is too coarse-grained. In order to capture the causal intuitions evoked by the example cases we need the concepts of CAC, PAC, and TAC.

I argue that the framework of causal models (and counterfactual conditionals more generally) provides us with the conceptual resources to distinguish a plurality of concepts of actual causation. Thus, my account is closely related to the internal pluralism that has been provided by Woodward in *Making Things Happen*. It is different from Woodward's account in that it extends pluralism to concepts of actual causation.

Since the distinctions that I suggest here are so closely related to the distinctions introduced by Woodward, it will be useful to look at them in some detail and highlight the differences. Woodward sets out to make the relation between manipulability and causation precise. A natural way to begin the analysis is to suppose that manipulability is both a necessary and a sufficient criterion for causation. This idea is captured by Woodward's definition of total cause:

> "X is a *total cause* of Y if and only if there is a possible intervention on X that will change Y or the probability of Y" (Woodward, 2003, 51).

However, there are counterexamples to manipulability being a necessary condition. Suppose we have a causal model consisting of three variables X, Y, and Z

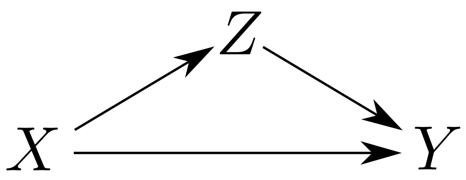

Figure 4.4.: Violation of faithfulness: the direct influence of X on Y is cancelled out by the indirect influence that X exerts on Y via the causal path going through Z.

whose dependence is described by the following structural equations:

$$Y = aX + cZ$$

$$Z = bX$$

where a, b, and c are fixed coefficients. Figure 4.4 displays the corresponding causal graph. Suppose that $a = -bc$. Then the direct causal influence of X on Y is cancelled out by the indirect causal influence that travels along the causal path going through Z. This causal model represents a case where a causal relation between X and Y exists but intervening only on X does not lead to a change in Y.

An example with this kind of causal structure is Hesslow's (1976) birth control pill case. Birth control pills (X) lower the probability of pregnancy (Z) and thereby lower the probability of thrombosis (Y), which can be a consequence of pregnancy. Besides that, birth control pills also have the side effect of increasing the probability of thrombosis. Suppose that the probability increase exactly cancels out the probability decrease that is mediated by the prevention of pregnancy (this is analogous to the assumption that $a = -bc$). Then the probability of thrombosis cannot be changed by an intervention that only targets birth control pills. Yet, at the same time, it is implausible to deny that birth control pills in some sense cause thrombosis.

Note that the birth-control case has a causal structure that is very similar to the structure of early preemption cases. In both kinds of cases there is one direct causal relationship that mediates a positive influence and another causal path that sustains the effect in case the cause is inactive (e.g. birth-control pills are not taken or the assassin in training is not pulling the trigger of his gun). In this sense one could say that taking birth-control pills *preempts* pregnancy as a cause of thrombosis.

The fact that Y (the probability of thrombosis) cannot be influenced by an intervention only on X (birth control pills) is reflected by the fact that X is not a total cause of Y. However, we may also be interested in what happens if we combine the intervention on X with an intervention that keeps Z fixed. Then the intervention on X *will* change the value of Y. For instance, if women take birth control pills even if they are already pregnant, then the pills will increase the probability of contracting thrombosis. Likewise, birth control pills lead to an increase of the probability of contracting thrombosis in women who cannot become pregnant for other reasons. This kind of causal dependence is captured by the notions of direct cause and contributing cause.

We address direct causes first. According to Woodward,

> "[a] necessary and sufficient condition 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" (55).

The thrombosis case describes a situation where just one other variable (Z, representing pregnancy) has to be kept fixed by an additional intervention such that the causal dependence of Y on X is revealed. The definition of direct cause generalizes this insight, stating that *all* other variables in $\mathcal{V}$[9] need to be held fixed.[10] Moreover,

[9] There is no difference between $\mathcal{V}$ and **V**. Both symbols refer to the set of variables that constitute the causal model.

[10] More precisely, it suffices to hold fixed at least one intermediate variable on each causal route from X to Y that is not the (supposedly) direct causal route. This will block causal influence that is transmitted along the corresponding causal route. Variables that do not lie on a causal route from X to Y do not need to be intervened upon.

the thrombosis case is a situation where the causal dependence of Y on X is revealed if Z is held fixed at *any* of its possible values. That is, birth control pills have adverse effects on women who are already pregnant *and* women who will not be pregnant for other reasons. For X to be a direct cause of Y it is sufficient that intervening on X leads to a change in Y when the other variables in $\mathcal{V}$ are held fixed at *some* value.

The notion of direct cause depends upon $\mathcal{V}$. For example, X is a direct cause of Y in the model given above. But suppose we interpolate a variable Z^* between X and Y, describing, for example, the dispersion of blood chemicals in the organism of a person who has taken birth control pills. Then X is no longer a direct cause of Y because the definition requires us to keep Z^* fixed. Yet, there is an important sense in which our considering birth control pills to be a cause of thrombosis is independent of whether the model includes such intermediate variables. Thus, we need a notion of cause that enables us to distinguish the influence along different causal routes independently of whether these routes are direct links between cause and effect or causal chains. This is captured by the notion of contributing cause which is defined as follows.

> "A necessary and sufficient condition for X to be a (type-level) *contributing cause* of Y with respect to variable set $\mathcal{V}$ is that (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship [...]; and that (ii) there be some intervention on X that will change Y when all other variables in $\mathcal{V}$ that are not on this path are fixed at some value" (59).

The first part (i) of the definition generalizes the definition of direct cause to the effect that it is no longer relevant whether X and Y are connected by a direct link or by a causal chain. The second part (ii) of the definition accounts for issues associated with failures of transitivity. Transitivity is violated, for example, in cases like "Dog Bite" (McDermott, 1995). In this scenario a right-handed terrorist plans to detonate a bomb. Yet, a dog bites off the terrorist's right forefinger. Therefore, the terrorist uses his left forefinger instead to detonate the bomb. Thus, the dog bite causes the terrorist's detonating the bomb with left forefinger and the terrorist's detonating the bomb with the left forefinger causes the explosion. Yet, we do not consider the dog bite to be a cause of the explosion. Woodward's condition (ii) excludes such cases.

According to Woodward, actual causation is another causal concept that needs to be distinguished from the concepts of total, direct, and contributing cause. However, he does not acknowledge that there are different concepts of actual causation. In fact, Woodward gives two definitions of actual causation (Woodward, 2003, 77 and 84). The first definition is along the lines of path-changing actual cause and the second definition is along the lines of contributing actual cause. However, he flags the first definition as a "first pass" notion (ibid., 77) and replaces it by the second notion once it is introduced.

Instead, I suggest the taxonomy given in table 4.1. In the left column there are Woodward's concepts of total and contributing cause (the concept of direct cause could be listed here as well). Woodward seems to suggest that actual causation is just one more entry in this column. Instead, I argue that actual causation is the heading of a whole new column of different causal concepts (and that, correspondingly, the other causal concepts described by Woodward are concepts of potential causation).

Potential Causation	**Actual Causation**
total cause	total actual cause
contributing cause	path-changing actual cause
	contributing actual cause

Table 4.1.: Internal causal pluralism with regard to potential and actual causation.

What is the relation between the concepts in the left column of table 4.1 and the corresponding notions in the right column? First, there is a difference in the relata. Potential causation describes relations between variables X and Y, while actual causation describes relations between variables that take on particular values ($X = x$ and $Y = y$).

Second, the notion of total causation is much more permissive than the notion of total actual causation. X is a *total cause* of Y in model M if there is *some* context $\vec{u}$, that is, some value assignment to the model's exogenous variables such that there is an intervention on X that leads to a change in Y. In particular, this context does not have to be the actual context or even a context that is likely to be instantiated. By contrast, $X = x$ is a *total actual cause* of $Y = y$ only if there is an intervention on X that leads to a change in Y in the *actual* context $\vec{u}$ (see also the discussion in section 1.2 for an argument why the more restrictive concept of actual causation is so important).

Finally, the pluralisms of potential and actual causation are not exactly parallel. What explains the fact that there is no equivalent to the distinction between path-changing and contributing actual causes on the side of potential causation? Path-changing actual causation entails holding fixed certain variables at actual values. Contributing actual causation entails holding fixed certain variables at non-actual values. But this distinction does not apply to definitions of potential cause because here all possible values (the actual and the non-actual) are treated on a par.

4.3. A Functional Justification

So far I have argued that distinguishing total, path-changing, and contributing actual causation accounts for the causal intuitions evoked by the three examples given in section 4.1. In this section I will provide an argument for the pluralist account that is based on the function that the concepts of actual causation have. The difference between total, path-changing, and contributing actual causes matters from the perspective of an intervening agent. In others words: conflating these different kinds of actual causation can impose a serious limitation on the practical value of the notion of actual causation. The pluralist account that I suggest here is thus not merely a reaction to the problems that extant unified accounts of actual causation face. Instead it is positively justified by considerations concerning why we should have concepts of actual causation in the first place.

Maybe intervening agents do not need to employ a philosophical theory of actual causation at all. In fact the police officers, electricians and policy makers whose perspectives we will consider in this section most likely have no explicit theory of actual causation. My point is: were these agents to employ a unified account of actual causation similar to the ones discussed in the previous chapter, then they would encounter problems. They are better off if they employ my pluralist theory.

Suppose I know that the assassin in training went on her mission alone, without his supervisor. If nothing interferes with his mission, then the trainee will pull the trigger of his gun and the victim will die. Suppose also that I am a police officer and wish to save the victim. In order to save the victim I have to intervene on the trainee's actions such that his mission fails. That is, a simple intervention on the total actual cause is sufficient in order to accomplish my goal.

Compare this with the situation where the assassin in training is a contributing actual cause, for example, because he is accompanied by another assassin who shoots at the victim at the same time and hits at the same time (symmetrical overdetermination). A simple intervention on the assassin in training will not save the victim. In order to save the victim, we have to intervene on *all* contributing causes. Thus, the overdetermination case describes a situation where the intervening agent fails to achieve her goal if she mistakenly identifies a contributing actual cause for a total actual cause. This is problematic. If we take concepts of actual causation to inform us about strategies to achieve our goals by means of intervention, they should at least be able to reflect this important difference.

In the assassin example there are only two contributing actual causes and both are straightforwardly identified as such. But of course the notion of contributing cause is much more general. It also applies to situations where the effect is multiply overdetermined such that an intervention on a contributing actual cause becomes effective only if it is combined with a large number of additional interventions. Voting scenarios are an example for this kind of situation (Chockler and Halpern, 2004; Livengood, 2013). Suppose you are supporting a policy that is submitted for vote to a board constituted of 11 members. Suppose also that the policy will be put in place if there is a simple majority for it. You also know that currently there is a majority of 8 members against your policy while it is supported by the remaining three board members. Each of the 8 opposing board members would turn out as contributing actual causes of your policy not being put in place. Convincing any one of the board member will not make a difference to the prospects of your policy. You will have to convince at least three members.

Now, aren't such a functional argument and pluralism about actual causation separate issues? If one takes causal pluralism to mean extramural pluralism, then this may well be the case. After all, I spell out total, path-changing, and contributing actual causation in one single theoretical framework: interventionist causal models. However, intramural pluralism is supported by the functional argument. The functional approach evaluates whether and to which degree causal concepts facilitate intervention. I agree with extant accounts that intervention is facilitated by considerations about actual causes. But I argue that intervention is facilitated even better if we distinguish total, path-changing, and contributing actual causation. Consequently, I argue, we should reject the assumption that there is a unified concept of actual causation and think in terms of total, path-changing, and contributing actual causation instead.

Earlier I have stated that we tend to identify both the major and the sergeant as actual causes of the corporal's advancing. Pace Lewis's and Schaffer's intuition, this judgement is justified from the interventionist perspective. Both the major and the sergeant are potential targets for intervention if we want to prevent the corporal from advancing. It will not be sufficient to stop either one of the major and the sergeant to give their orders and to make them do nothing instead. We need to intervene on both to prevent the outcome. In this sense the trumping case is just like a situation of symmetrical overdetermination, where two contributing actual causes bring about an effect.

Unlike cases of symmetrical overdetermination, trumping also involves an asymmetry between the major and the sergeant, which is related to the fact that only the major is also a total actual cause. Again, this can be explained from the interventionist perspective. There is a straightforward sense in which the total actual cause is a better target for intervention: whereas an intervention on a mere contributing actual cause needs to be combined with other interventions, intervening on a total actual cause allows direct control over the outcome.

I conclude that from an interventionist perspective the ability to draw a conceptual difference between total actual causes and contributing actual causes is essential for accomplishing goals. We shall now turn to path-changing actual causes.

Path-changing actual causes are a subclass of contributing actual causes. Again, we need (at least) two interventions in order to avoid the outcome. The important difference between path-changing actual causes (PACs) and contributing actual causes that are not path-changing actual causes (CAC\PACs) is the relation between the interventions that have to be applied. In the case of CAC\PACs we need to apply two interventions that target independently active causal processes. In the case of PACs we need to combine a primary intervention that targets the path-changing actual cause with a secondary intervention that eliminates a threat to the goal that did not exist (or at least was not as acute) before the primary intervention was applied. That is, the secondary intervention needs to be performed in order to counteract the adverse consequences of the primary intervention.

For example, in "Backup," where the assassin in training is a path-changing actual cause of the victim's death, we need to apply two interventions. We need to apply a primary intervention on the assassin in training. But we also need to apply a secondary intervention on the supervising assassin because otherwise the supervisor would attack the victim. The difference to the overdetermination case is that the supervisor would attack the victim only as a result of the primary intervention.

This is an important difference for epistemic reasons. Elements of a set of CAC\PACs are more straightforwardly identified as threats to the desired outcome because each of them corresponds to an active causal process. Situations involving path-changing actual causes can be much more difficult to handle because sometimes we find out about the backup threats only through applying the primary intervention. And this is because the backups are only activated as a result of the primary intervention.

This is even more problematic considering the fact that there are cases where intervening only on the path-changing cause without applying a secondary intervention that counteracts the adverse consequences can make the situation even worse. Suppose, for example, that the assassin in training is known to hurt his victims severely but does not kill them. His supervisor, however, always hits her victims lethally. Suppose also that, given there is no interference, the trainee's attack preempts the supervisor's attack such that the victim is only hurt but not hit lethally. Intervening on the assassin in training's mission (but not on the supervisor) would have detrimental consequences to our goal of saving the victim. As a result of our intervention the supervisor will kill the victim.

There are other examples with this kind of structure. First, consider safety mechanisms. Suppose, for example, you want to eat toast. However, as you put the bread into the toaster the fuse is blown. You stick the fuse back in but it keeps being blown whenever you start the toaster. You think you should intervene on the faulty fuse because you identify it as the actual cause of your not being able to enjoy your toast. However, the fuse is only a path-changing actual cause. It preempts a short-circuit and fire which would be an actual cause of your not enjoying the toast—and other consequences that may be much worse—if you disabled the fuse.

Second, consider complex policy decisions. Suppose the members of a city council want to make cycling safer and consider issuing a law that requires cyclists to wear a helmet.[11] Helmets are generally considered to increase cycling safety. However, wearing a helmet can also negatively affect safety because motorists are encouraged to leave less space when overtaking cyclists wearing a helmet (Walker (2007)). Suppose that Suzy is a cyclist. Before the bicycle law was issued Suzy used to cycle without a helmet. But now that the law is in place she wears a helmet while riding. A car overtakes while leaving little space, Suzy falls and due to her wearing a helmet she does not suffer head injuries. We are tempted to say that Suzy benefited from the helmet law because her wearing the helmet prevented severe head injuries. However, whether this is true depends on whether not wearing a helmet is a total actual cause or only a path-changing actual cause of an increased risk during cycling. If not wearing a bicycle helmet is a total actual cause, then Suzy seems to have benefited from the law. But if not wearing a helmet is only a path-changing actual cause then this is not necessarily so. It could be the case that without the helmet the motorist would have left more space and Suzy had not fallen in the first place.

I conclude that total, path-changing, and contributing actual causation are three

[11] This example is taken from Hitchcock (2017).

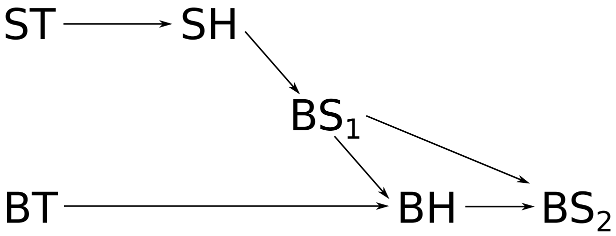

Figure 4.5.: The revised causal model of the Suzy-Billy case.

concepts of actual causation that should be distinguished for reasons associated with the implications of these concepts in particular contexts of intervention.

Before we turn to the next section note the following. So far I have discussed instances of early preemption. Do the same arguments apply to late preemption? Figure 4.5 shows a simplified version of Halpern and Pearl's (2005) time-indexed model of the Suzy-Billy case that we have discussed in Chapter 3. According to this model, Suzy's hitting the bottle $SH = 1$ is a cause of the bottle's shattering $BS_1 = 1$. The bottle's being shattered then causes Billy's missing the bottle $BH = 0$. The bottle's being shattered at the final stage is governed by the structural equation $BS_2 = BS_1 \vee BH$. In the actual situation the bottle is shattered at the final stage ($BS_2 = 1$) because it was already shattered at the earlier stage $BS_1 = 1$.

Suppose it is our goal that the bottle be intact at the final stage ($BS_2 = 0$). Then we need to combine a primary intervention on Suzy's throwing the stone with a secondary intervention on BH. As in cases of early preemption this secondary intervention can be described as an intervention that needs to be applied in order to prevent Billy's hitting the bottle which would otherwise be an adverse consequence of our primary intervention.

However, there is also a sense in which it seems to be wrong to describe Billy's hitting the bottle as an adverse consequence of the primary intervention. After all, the only reason for Billy's failure to hit the bottle in the first place is the fact that the bottle is already broken at the time Billy's stone arrives. In fact, Billy's throwing his stone is an alternative causal process that represents an independent threat to the bottle (as long as the bottle is not destroyed). One consequence of this is that the epistemic difficulties that are characteristic of situations involving early preemption do not seem to apply to cases of late preemption.

This means that the functional theory that I am proposing here seems to characterize the Suzy-Billy case as a case that is much more similar to instances of symmetrical overdetermination, which also require applying two independent interventions. But if this is so how do we explain the fact that in the Suzy-Billy case we identify only Suzy as actual cause whereas in cases of symmetrical overdetermination we identify both factors as actual causes? I will discuss this problem in more detail in Chapter 5. The point is that cases of late preemption indicate a limitation of my interventionist account of the function of actual causation. But they also indicate a limitation of interventionist accounts more generally. Instead, I will argue, we need to refer to responsibility in order to explain our causal intuitions in such cases.

4.4. Two Kinds of Context-Sensitivity

What is the place of norm-dependence within the pluralist account of actual causation presented here? In Chapter 3 I have discussed Halpern and Hitchcock's (2015) approach according to which norm-dependent considerations enter through an additional requirement on the witness world of the actual cause. More specifically, this approach requires that the witness $s_{\vec{X}=\vec{x}',\vec{W}=\vec{w}',\vec{u}}$ of $\vec{X}$ being an actual cause be at least as normal as the actual world $s_{\vec{u}}$. In principle, this requirement could be added to each of the definitions of total, path-changing, and contributing actual causation such that we arrive at norm-dependent versions of these definitions.

However, in this section I will have a closer look at how exactly these different concepts of actual causation interact with Halpern and Hitchcock's requirement. Halpern and Hitchcock's idea of employing a normality ranking over possible worlds treats all variables of the model on a par. In particular, it does not distinguish between the specific normality considerations that concern the purported actual cause $X = x$, on the one hand, and variables $\vec{W} = \vec{w}$, on the other hand. But this should be surprising because the kind of information that we have about default behaviour can be quite different. The typicality considerations with regard to $X = x$ are concerned with whether X typically takes on value x and whether there are other values x' that are more typical. The typicality considerations with regard to $\vec{W} = \vec{w}$ depend on similar considerations. But in addition to that our expectations regarding the values of those variables in $\vec{W}$ that depend on X should be influenced by the information provided by the structural equations and the values of their parent variables.

In order to illustrate the point let us consider "Backup," where we take the assassin in training to be a path-changing actual cause of the victim's death. The corresponding witness is the world where the trainee does not shoot, where we keep fixed that the supervisor does not shoot, and where the victim does not die ($T = 0, S = 0, VD = 0$). We need to show that the witness is at least as normal as the actual world ($T = 1, S = 0, VD = 1$). There are two kinds of context-sensitive considerations that feed into this evaluation.

First, there is the context-sensitivity of considerations about the default values of individual variables (context-sensitivity$_1$). These considerations depend upon the context to the degree that it depends on the context what value we expect a variable to take on or what value a variable should take on. Typically, we think that assassinating a person is morally wrong, against the law, and unusual. But there may be contexts where this is different, for example, if the victim is a criminal that threatens to kill a group of innocent hostages and there is no other way to rescue them.

Second, there is a kind of context-sensitivity that concerns potential violations of structural equations (context-sensitivity$_2$). According to the structural equations, the supervising assassin kills the victim if the assassin in training fails to do so ($S = \neg T$). In that sense, it seems that the witness describes a situation that is highly *abnormal*. Again, the degree to which this is to be taken as an abnormal situation depends on the context. For example, there may be reasons to believe that the supervising assassin will dismiss her orders or loose her nerves when pulling the trigger of her gun.

How are these two kinds of context-sensitive considerations to be related to each other? This is particularly interesting in cases where the corresponding typicality considerations pull in different directions. "Backup" is such a situation. In most contexts we think that suitable defaults are $T = 0$, $S = 0$, and $VD = 0$. This indicates that the witness is more normal than the actual world. At the same time, however, the witness represents a situation where the structural equation $S = \neg T$ is violated, suggesting that the witness is less normal than the actual world (where no structural equation is violated). In "Backup" the considerations regarding the individual values seem to dominate because there is a clear intuition according to which the trainee is an actual cause of the victim's death (and at least according to Halpern and Hitchcock's account this would not be the case if the witness were not at least as normal as the actual world). In analogy to Halpern and Hitchcock's (2015) treatment of the "Careful Poisoning" case discussed in section 3.8 it seems to be the case that a plausible ranking for possible values of S is the following (with decreasing order of normality):

(1) $S = 0$, regardless of T

(2) $S = \neg T$

(3) $S = 1$, regardless of T

That is, the most normal case is the case where the supervisor does not intend to kill the victim independently of the trainee's actions. Second on the ranking is the case where the supervisor sticks to her orders and kills the victim if the trainee fails to do so. The least normal case is where the supervisor intends to attack the victim regardless of the trainee's actions.

Yet, going back to the discussions of early preemption and transitivity in section 3.3, there are also instances where considerations regarding the violation of structural equations seem to dominate. Consider the boulder case: a boulder is dislodged, the hiker sees it, ducks, and survives (figure 4.6). The boulder is an actual cause of the hiker's ducking, the hiker's ducking is an actual cause of the hiker's survival, but clearly the boulder's being dislodged is not an actual cause of the hiker's survival. According to our definition of total actual cause, this is the case because there is no counterfactual dependence of the hiker's survival on the boulder's being dislodged. Neither is the boulder's being dislodged a path-changing or contributing actual cause because there is no other variable that could be held fixed at an actual or non-actual value such that the counterfactual dependence would be restored.

In section 3.3 we have seen that this is a result of a particular model choice. If we include a variable B representing the boulder's presence close to the hiker's head such that the hiker has no opportunity to save herself, then we could restore a counterfactual dependence, given that we hold fixed the value of B. As a result, we would get the undesirable outcome that the boulder's being dislodged *is* a path-changing actual cause of the hiker's survival.

Hitchcock (2001), whose active route account faces exactly this problem, argues that the causal model that includes variable B is not an apt model. This is because it represents the far-fetched scenario where a boulder appears close to the hiker's head

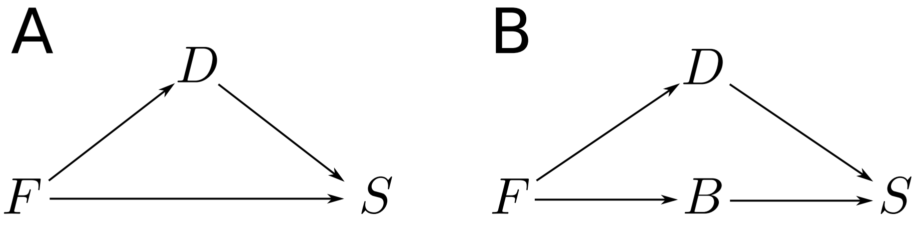

Figure 4.6.: Boulder. According to model A, the boulder's being dislodged is not an actual cause of the hiker's survival, which is the intuitive judgement. According to model B, however, the boulder's being dislodged *is* an actual cause of the hiker's survival. But this result can be rejected either by showing that model B is a non-apt model or by incorporating considerations of normality. The structural equations of model B are $F = 1$, $D = F$, $B = F$, and $S = \neg B \vee D$. See section 3.3 for a detailed discussion.

even though it was never dislodged. I agree that this is a legitimate treatment of the case. But let us consider what happens if we include the variable nevertheless and apply the normality criterion.[12] In the actual world the boulder falls and approaches the hiker's head, the hiker ducks, and survives ($F = 1$, $B = 1$, $D = 1$, $S = 1$). The witness of the boulder's falling being an actual cause of the hiker's survival is the world where the boulder is not dislodged but appears close to the hiker's head, nevertheless, the hiker does not duck, and the hiker dies ($F = 0, B = 1, D = 0, S = 0$). How do these worlds compare in terms of normality? Looking at the individual variables, the result is not clear. The fact that in the witness world the hiker dies seems to indicate that this is a world that is more abnormal than the actual world (yet, sometimes hikers die). The fact that no boulder is falling in the witness world could speak for this world being more normal (most of the time boulders do not fall). But looking at the *combination* of variables clearly decides the comparison: the boulder that appears close to the hiker's head without ever having fallen clearly makes this a highly abnormal scenario. Thus, I conclude that restrictions that structural equations impose on considerations of normality sometimes dominate

[12] In Chapter 7 I will provide a more detailed analysis of the relation between constraints of apt modelling and the relevance of the default/deviant distinction.

those considerations related to the typicality of individual variables.

4.5. Conclusion

In this chapter I have advanced a pluralist account with regard to the notion of actual causation. In support of the account I have provided two lines of argument. First, I have presented three cases that raise difficulties for unified accounts of actual causation and I have argued that distinguishing between total, path-changing, and contributing actual causation helps account for these cases. Second, I have provided a functional argument for distinguishing between total, path-changing, and contributing actual causation. The distinction between total and contributing actual causation is important because in cases involving total actual causation only one intervention is required in order to prevent the effect whereas in cases involving contributing actual causation (but not total actual causation) more than one intervention is required. Cases involving path-changing actual causation also require more than one intervention. Moreover, these cases are different from cases involving contributing actual causation (but not path-changing actual causation) because we need to distinguish between primary interventions that target the path-changing actual cause and secondary interventions that counteract the adverse consequences of the primary intervention. Finally, I have distinguished two kinds of context-sensitivity. Context-sensitivity$_1$ concerns considerations regarding the default values of individual variables. Context-sensitivity$_2$ concerns considerations regarding combinations of variables and values that reflect a violation of the structural equations. These different kinds of context-sensitivity will be addressed in more detail in Chapter 6, where I will discuss the concept of actual causation in the law. But before that we shall address a key assumption of the functional argument provided in this chapter: the assumption that concepts of actual causation facilitate the purpose of intervention. This will be done in the following chapter.

5. Responsibility and the Limits of Interventionism

A key assumption of the foregoing chapter has been that concepts of actual causation have an important role to play where agents are interested in manipulating the outcome of a particular situation. With this assumption I have taken up the interventionist tradition which is dominant in the literature on causal models and actual causation. *Prima facie* the focus on intervention should come as a surprise because the concept of actual causation has been imported from the law, where its function is to facilitate the *post hoc* assessment of responsibility. Most contributors to the interventionist literature on actual causation seem to be aware of this.[1] Thus, if they attempt a clarification of the concept with reference to intervention, they seem to assume that responsibility and intervention are closely related.

In this chapter I shall have a closer look at the relation between intervention and responsibility with particular regard to the debate on actual causation. In the recent literature there has been a debate about how selective causal attributions as evoked by norm-violating behaviour are to be interpreted. Interventionists (Hitchcock and Knobe, 2009) argue that causal reasoners legitimately select norm-violating behaviour because such behaviour is particularly suited as a target for corrective intervention. By contrast, proponents of the responsibility view (Sytsma et al., 2012)

[1] For example, Pearl (2000), Halpern and Pearl (2005), and Hitchcock (2017) explicitly identify 'actual causation' as a notion that is relevant for ascribing responsibility.

argue that causal reasoners legitimately select norm-violating behaviour because such behaviour indicates responsibility for the outcome.[2] In this chapter I will examine whether and in which way these accounts are distinct accounts in the first place. According to Hitchcock and Knobe, ascribing responsibility to an agent can be understood as one specific kind of intervention. Blaming an agent for their undesirable behaviour then amounts to an intervention on the agent's motives such that the undesirable behaviour is discouraged. Thus, it seems intervention and responsibility do not provide explanations that compete with each other. Instead the interventionist account seems to be simply the more encompassing theory.

However, the idea that practices of ascribing responsibility can be understood as a particular form of corrective intervention presupposes a consequentialist view of responsibility that is not uncontroversial. I will contrast this view with a retributivist account, according to which our practices of ascribing responsibility are not justified through their consequences but trough a backward-looking assessment of desert. Employing the distinctions drawn in the foregoing chapter I will then discuss two instances of reasoning with concepts of actual causation where the interventionist account faces limitations and reference to a retributivist notion of responsibility gives a better explanation of our causal intuitions. First, total actual causes are typically norm-violating factors and they are distinguished from background conditions that are typically norm-conforming. According to the interventionist account, these norm-violating factors are particularly suited as targets for intervention. However, there are instances where such norm-violating factors should not be intervened upon. Second, there are cases of late preemption where it is not clear what practical inferences an intervening agent should draw from claims of path-changing actual causation. In these cases it is not clear in what sense the path-changing actual cause is more suited as a target than other factors that are not

[2] There is a third view, based on the Culpable Control Model (Alicke et al., 2011), according to which the selective judgement is a bias that results from a desire to blame.

actual causes.

In both kinds of cases the relevant causal claims are better explained by our interest in the assignment of retributive responsibility. Practices of blaming and praising that follow a retributive ideal derive their justification from the past wrongs and goods that they respond to. Retributive responsibility makes sense of selective claims of total actual causation that concern the past, even if those factors are not suitable targets for intervention. Retributive responsibility also explains why it matters to point out path-changing actual causes in cases of late preemption. The reason why these causal claims are better explained from the perspective of retributive responsibility is related to a difference in perspective. Retributive responsibility evaluates causes in a retrospective way whereas the interventionist account evaluates causes in terms of their usefulness for future action.

What can we learn from this? Functional accounts of causation seem to depend on two questions. First, given that a concept has a particular function, what can we infer about the concept? Second, given a certain concept, what is its function? In the foregoing chapter I focused on the first question, arguing that from an interventionist perspective we should be pluralists with regard to actual causation. In the Introduction I also argued that a functional account along these lines can be understood as an instance of conceptual engineering that aims at revising our causal concepts. A concept like the concept of actual causation (or a set of concepts) can be seen as a tool for achieving certain goals. And if there are ways to improve the tool such that the goal is achieved in a more effective way, then there may be good reasons to do so. In this chapter I address the second question: what is the function of concepts of actual causation? In particular, I will suggest that the concepts fulfil more than one function. The lesson is that a functional account has to be cautious if it aims at revision: a concept may serve more than one purpose and revising the concept such that it facilitates this one purpose best may come at the cost of neglecting other purposes.

Let me briefly clarify what this chapter is *not* trying to achieve. First, I do not aim to give a reductive account of causation, that is, an account that tells us what causation *is* in terms of an underlying notion of responsibility. I do not think that such an account would be illuminating because I take the notion of responsibility itself to be dependent upon an underlying notion of causation. And, thus, a reductive theory would turn out to be circular.[3]

Second, in this chapter I ask: given certain causal intuitions what function do they serve? That is, I will assume that we have certain causal intuitions and give a functional account that explains these intuitions. In particular, this chapter does not attempt to justify these intuitions. A justification of the relevant causal intuitions would require a defence of the corresponding retributivist theory of responsibility. I do not aim to give such a defence. Instead I will argue that *if* retributivists give an adequate description of our actual practices of assigning responsibility, then they also account for certain causal intuitions.

Third, by saying that incorporating considerations of responsibility can provide a better explanation of certain causal intuitions I do not mean to provide a functional account that is complete. Claims of actual causation may facilitate functions beyond intervention and responsibility, such as explanation and prediction.

The chapter is structured as follows. In section 5.1 I will briefly review the main competitors in the current debate about the function of actual causation. In

[3] In fact, the prospectives of a reductive theory in terms of responsibility may be even worse than the prospectives of an interventionist theory that would attempt to be reductive. Woodward (2003), for example, interdefines causation and intervention which raises an issue of circularity. However, according to Woodward, this is not to be seen as problematic because defining the causal relation between two variables X and Y in terms of interventions requires us to assume causal relations *other than* the one holding between X and Y. In particular, we need to assume that the intervention variable I is a cause of X, that it acts as a switch on X (interrupting all other influences on X), that there is no influence of I on Y that is not mediated by X. And we need to make sure that I is not correlated with Y in a way that is not mediated by X. This response, however, is not available to the proponent of an analogous responsibilist theory of causation. If we were to define causation in terms of responsibility, then agent A's causing effect E would be defined through A's being responsible for E.

section 5.2 I will introduce the idea that interventionism is not opposed to the responsibility view but simply a more encompassing theory. In section 5.3 I will turn to responsibility and introduce two kinds of competing accounts. First, I will discuss consequentialist accounts of responsibility according to which our practices of assigning responsibility essentially derive their justification from their positive effects on future behaviour. Second, I will discuss retributivist accounts of responsibility according to which responsibility is assigned to those who deserve it. I will then discuss two kinds of reasoning with concepts of actual causation where the interventionist account faces limitations and reference to a retributivist notion of responsibility gives a better explanation of our causal intuitions. In section 5.4 I will address issues related to selective claims of total actual causation. In section 5.5 I will turn to cases involving late preemption.

5.1. Intervention, Responsibility, and Blame

Much of the recent debate about the function of actual causation arises from studies like the one involving the pen vignette (see Chapter 2). The pen vignette describes a situation where the department's receptionist has no pens because Professor Smith and an administrative assistant both took one of the last two remaining pens. The receptionist's having no pens depends symmetrically on both the professor's and the administrator's taking a pen. But test subjects typically identify Professor Smith as the actual cause of the receptionist's problem. And this is thought to be related to the fact that only the administrative assistants are allowed to take pens while members of faculty have to buy their own.

Broadly speaking, there are three kinds of explanations that have been provided for this result. First, according to the interventionist position defended by Hitchcock and Knobe (2009), the test subjects consider two kinds of counterfactual scenarios:

one where the receptionist's problem does not occur because the administrator acts differently and one where the problem does not occur because Professor Smith acts differently. Moreover, test subjects tend to identify the counterfactual involving a change in Professor Smith's behaviour as the more relevant scenario because it is one that is more normal. Besides the descriptive claim that we typically tend to identify norm-violating factors as relevant causes, Hitchcock and Knobe also make an evaluative claim. They argue that a focus on norm-violating events is justified because these events correspond to factors that are particularly suited as targets for intervention. The underlying reasoning depends on the kind of norm that is being violated. First, it is reasonable to intervene on factors that violate *statistical norms* because such interventions amount to generalizable strategies. Second, it is reasonable to intervene on factors that violate *moral norms* because such interventions increase the overall number of morally good aspects of a situation. Third, it is reasonable to intervene on factors that violate *functional norms* because such interventions increase the overall functionality of the system that is being intervened upon. In the following sections we will discuss these justifications in more detail, but here is an initial example of how this kind of reasoning applies to a particular case: Professor Smith seems to be a suitable target for intervention, presumably, because intervening on Professor Smith reinforces a policy that is already in place. By contrast, intervening on the assistant would undermine the current policy, by way of encouraging other faculty members not to comply with it or discouraging other assistants to make use of their privilege.

What kind of norm does Professor Smith violate? Hitchcock and Knobe (2009, 608f) suggest this as an instance where a moral norm is violated. However, there does not seem to be anything morally bad about Professor Smith's taking the pen, other than that it undermines an agreed upon rule. Instead, it seems that the norm can be described as a functional norm, that is, a norm that ensure that administrative processes in the department run smoothly.

Second, Sytsma, Livengood, and Rose have suggested an interpretation of the results in terms of responsibility. According to the "responsibility view," as defended by these authors, the pen vignette tracks an "ordinary concept of causation" that is an "inherently normative concept: Causal attributions are typically used to indicate something more akin to who is *responsible* for a given outcome than who caused the outcome in the descriptive sense of the term [...]" (2012, 815). The difference to the interventionist explanation that will be most relevant in the following is that the causal attributions are explained with regard to the role that they play in assigning responsibility. Interestingly, the authors do not specify what responsibility is. However, the discussion in section 5.3 will show that distinguishing, for example, consequentialist from retrubitivist views has important ramifications.

Sytsma et al. support the responsibility view with a more detailed analysis of the role that statistical norms (or typicality) play in selective causal judgement. As we have seen in Chapter 2, they argue that we need to distinguish two types of typicality in order to characterize the role of statistical norms in causal judgement. First, there is typicality that relates to "how *people* generally behave in a given type of situation" (2012, 816). This is called population-level typicality. Second, there is typicality that relates to how a particular agent "herself generally behaves in [a given type of] situation" (ibid.). This is called agent-level typicality. Sytsma et al. provide evidence to the effect that, first, information about population-level typicality does not influence the test subjects' judgement with regard to the pen vignette. Second, they show that information about agent-level typicality does influence the test subjects' judgements. But it does so in the reverse way than commonly acknowledged. That is, an agent's causal role is emphasized if she acts agent-level typically rather than agent level atypically.

Sytsma et al. take these results to be predicted by their responsibility view. First, with regard to population-level typicality, they argue "that how other people typically act in a given type of situation will largely be treated as irrelevant to whether or not a specific person is taken to be normatively responsible for an outcome" (2012, 816). That is, the degree to which we take Professor Smith to be responsible for the problem is independent of whether other members of faculty take pens. Sytsma et al. concede that "excuses of the form "everybody was doing it" might help to explain an agent's action" but they suggest that "people generally do not take such excuses to actually mitigate normative responsibility" (ibid.). Second, with regard to agent-level typicality, Sytsma et al. argue that agents who habitually act in a way that has potentially bad consequences are rated to be more responsible than agents who do so only occasionally. The reason is that such habitual acts increase "the chance that the bad outcome would eventually occur" (ibid.). That is, specifying that Professor Smith regularly takes pens should increase her perceived responsibility for the receptionist's problem because the more often Professor Smith takes a pen, the more likely it is that the receptionist's problem occurs as a result of her actions.

Finally, according to the Culpable Control Model (henceforth CCM, Alicke (1992); Alicke et al. (2011)), an observer of some negative state of affairs first forms an initial blame hypothesis regarding the involved agent. The blame hypothesis is associated with an "active desire to blame" the agent and "[t]his desire, in turn, leads observers to interpret the available evidence in a way that supports their blame hypothesis" (Alicke et al., 2011, 675). For example, the initial blame hypothesis can lead to overrating the control that the agent has over her own behaviour or the outcome of her acts. The CCM predicts that the *outcome* of norm-violating behaviour has an influence on our blame judgements and our causal judgements. Here it differs from the interventionist account and the responsibility view which predict a dependence on whether the behaviour itself is a norm violation or not (independent of the outcome). Another important difference is that the CCM characterizes the difference in causal attributions in situations like the pen case as a bias. According to the interventionist and the responsibility account, selective causal judgement is justified.

5.2. Intervention vs Responsibility?

In the following I am primarily interested in the relation between the interventionist account and the responsibility view. In particular, I am interested in a claim put forward by Hitchcock and Knobe that concerns the function of claims of actual causation. Considering norm violation as a criterion for causal selection, they admit that "it is natural to assume that the purpose of this mechanism must have something to do with picking out the agents who are truly *to blame* for an outcome" (2009, 606, emphasis original). But they argue that this natural assumption can be identified as one special instance where the broader interventionist theory is at work: "One can regard the act of blaming a person as one way of intervening on that person's behavior and trying to get him or her to change" (ibid). That is, according to Hitchcock and Knobe, the interventionist view and the responsibility view are not theories that stand in opposition. Instead the responsibility view is to be seen as being incorporated into the interventionist theory. It represents the special case that involves intervention on human behaviour by means of ascribing responsibility.[4]

If this is true, then the broader interventionist account should be able to explain the evidence that Sytsma et al. provide in support of the responsibility view. At first sight, this does not seem to be the case. Hitchcock and Knobe argue that from the interventionist perspective it is reasonable to highlight those aspects of a situation

[4] Hitchcock and Knobe put this claim forward in order to address the CCM as a competitor. But the point, presumably, extends to Sytsma et al. 's responsibility view. Blaming corresponds to a form of intervention that discourages undesirable behaviour. Praising corresponds to a form of intervention that encourages desirable behaviour.

that violate statistical norms. But we have seen that statistical typicality is either irrelevant (if it concerns the population level) or that we highlight the statistically typical rather than the statistically atypical (if it concerns the agent level).

But let us have a closer look at Hitchcock and Knobe's explanation for why statistically atypical behaviour should be relevant from the perspective of intervention. Hitchcock and Knobe consider a situation where a scientific article is not accepted because one of the involved referees has the idiosyncratic view that an article should not use the word 'and' more than three times per page. Hitchcock and Knobe discuss two kinds of strategies for dealing with this situation. One strategy is to make sure that the paper is sent to a reviewer who does not employ this rule. This would amount to intervening on the aspect of the situation that is currently abnormal. Alternatively, one could allow the paper to be sent to the same reviewer and try to compensate the reviewer's abnormal criteria by introducing another abnormality: by using the word 'and' only three times per page. Hitchcock and Knobe argue that the first strategy is better because it is generalizable. There are many more situations where a paper will be accepted if it is sent to a reviewer without this idiosyncratic view than there are situations where a paper will be accepted because it uses the word 'and' only three times per page. Thus, the idea is that we should focus on the statistically abnormal because intervening on the statistically abnormal amounts to strategies that are generalizable.

Going back to Sytsma et al.'s results, it seems like the interventionist can explain the results after all. First, consider the result that an agent who acts agent-level typically in taking a pen is rated higher than an agent who acts agent-level atypically. From the interventionist perspective this makes sense because intervening on the agent who acts agent-level typically is a strategy that is more generalizable. There are many more situations where a lack of pens will be avoided if we intervene on those who systematically take the pens rather than on those who do so only occasionally. Interestingly, Sytsma et al. found that the information about agent-level typicality even overrides information about permissibility. An agent who acts against the policy but only occasionally is rated lower than an agent who conforms with the policy but acts agent-level typically in taking the pens. This indicates that the generalizability of the strategy seems to be even more important than whether a certain prescriptive norm is violated or not—a conclusion that seems to speak in favour of the interventionist perspective, rather than the responsibility view.

Second, consider the result that causal attribution is indifferent to population-level atypicality. This can be explained by the interventionist perspective as well. Whether an intervention on an agent is a generalizable strategy for solving the pen problem is independent of whether the agent belongs to a population that typically acts in this way or not (given that population-level typicality does not imply agent-level typicality). That is, whether Professor Smith is an individual that belongs to a group that typically does not take pens should have no consequences for interventions on Professor Smith. Population-level atypicality may have ramifications for interventions if interventions are performed on the population level, though. An example of such an intervention would be a reminder about the pen policy that is sent to all members of faculty rather than merely to Professor Smith. But this seems to be irrelevant in this particular case because the test subjects evaluate the causal relevance of particular individuals, not of whole populations.

Thus, it seems like Hitchcock and Knobe are right and interventionism is the more encompassing account. However, note that the consequentialist view of responsibility presumed by Hitchcock and Knobe is at least not uncontroversial, which they acknowledge (see Hitchcock and Knobe (2009), 606). Unfortunately, the proponents of the responsibility view do not specify what theory of responsibility they rely on. Thus, it should be useful to have a closer look at concepts of responsibility, which I will do in the following section.

5.3. Disambiguating Responsibility

Much of the philosophical literature on responsibility is concerned with the question whether responsible agency is compatible with causal determinism. Incompatibilists argue that responsible agency requires free will and that there is no free will if our actions are causally determined. Compatibilists, by contrast, argue that responsible agency and free will is possible in a causally determined world. However, our *practices* of assigning responsibility can be described independent of this problem, a point famously made by Strawson (1962). These practices of assigning responsibility shall matter in the following. A natural place to examine these practices is the law and, in particular, theories of punishment.

The first thing to note is that the word "responsibility" has a range of different senses. This is illustrated by Hart's (1968, 211) story about a captain, X:

> "As captain of the ship, X was responsible (1) for the safety of his passengers and crew. But on his last voyage he got drunk every night and was responsible (2) for the loss of the ship with all aboard. It was rumoured that he was insane, but the doctors considered that he was responsible (3) for his actions. Throughout the voyage he behaved quite irresponsibly (4a), and various incidents in his career showed that he was not a responsible (4b) person. He always maintained that the exceptional winter storms were responsible (5) for the loss of the ship, but in the legal proceedings brought against him he was found criminally responsible (6) for his negligent conduct, and in separate civil proceedings he was held legally responsible (7) for the loss of life and property. He is still alive and he is morally responsible (8) for the deaths of many women and children" (numbers added).

There are eight different appearances of the word 'responsible' and its cognates in the story. Here is an overview:[5]

(1) Role responsibility describes certain duties that a person has by occupying a particular place in social organization.

(2) Outcome responsibility describes a form of responsibility which arises from the effects of a person's actions and for which the person deserves praise or blame.

(3) Capacity responsibility describes whether a person is capable to be the author of their own actions.

(4) Virtue responsibility describes a person's character, reputation, intentions, or actions as dependable.

(5) Causal responsibility as in "the exceptional winter storms were responsible for the loss of the ship" describes instances where the expression 'responsible for' is synonymous to 'caused' or 'produced.'

(6) Criminal responsibility is a kind of liability that is determined by legal proceedings.

(7) Legal responsibility describes a kind of liability that is determined by civil proceedings.

(8) Moral responsibility describes responsibility that arises from the violation of moral norms.

These eight different senses of responsibility are highly interdependent and much could be said about the relations between them. However, these details shall not

[5] Hart distinguishes four main kinds of responsibility: role responsibility, causal responsibility, legal liability-responsibility and moral liability-responsibility. He focuses on these because he makes a point about the difference between legal and moral liability-responsibility. In this overview I draw from more recent taxonomies that have been provided by Cane (2002) and Vincent (2011).

matter in the following, where I will focus on a loose understanding of outcome responsibility. I take the term from Vincent (2011), who takes it from Perry (2000). According to Vincent, the advantage of this term is that it captures among other ideas the "idea of a form of responsibility which looks backwards in time to states of affairs (outcomes or actions) that occurred in the past" (2011, 17). I agree that considerations of outcome responsibility arise most naturally from outcomes that occurred in the past. Yet, it seems to be reasonable to extend the concept such that it also applies to outcomes that may occur in the future, given that an agent acts in a certain way. In fact, outcome responsibility seems to be better described as involving effect backward-reasoning which is also a feature of actual causation (see the discussion in Chapter 1). Moreover, I take outcome responsibility to be an important factor in assessing moral responsibility and at least in certain kinds of criminal and legal responsibility. Outcome responsibility is different from mere causal responsibility in that it is only assigned to agents and their actions, but not to objects. Finally, the assumption that outcome responsibility is backward-looking seems to distinguish it from role, virtue, and capacity responsibility, which are forward-looking concepts.

What are our practices of assigning outcome responsibility and how are they justified? In the following I shall look at two opposing strands in the literature. First, there are moral influence theories that justify practices of assigning responsibility by emphasizing the beneficial consequences of these practices. Second, there are retributivist theories according to which assigning responsibility is justified intrinsically.

Moral influence theories have roots in the empiricist ethics of Hobbes and Hume and have found their first prominent defence in Moritz Schlick's *Problems of Ethics* (1939). According to Schlick, the aim of imputing responsibility to a person is punishment (or reward). Punishment, in turn, "is an educative measure, and as such is

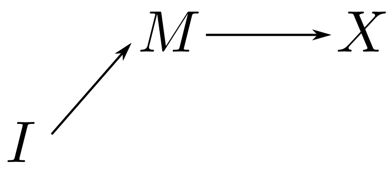

Figure 5.1.: Ascription of responsibility as intervention on the agent's motives.

a means to the formation of motives, which are in part to prevent the wrongdoer from repeating the act (reformation) and in part to prevent others from committing a similar act (intimidation)" (152). Schlick emphasizes that the question of responsibility, thus, literally is "the question concerning the *correct point of application of the motive*" and that "in this its meaning is completely exhausted" (153).[6]

Moral influence theorists, thus, construe the question of responsibility essentially as a question of efficient corrective intervention. One plausible way to spell out the analogy is to presume a causal model (figure 5.1) where an agent's action X depends causally on certain motives M. Blaming the agent for performing X then amounts to introducing measures I that change M and thus lead to a change in the agent's actions X. Interestingly, it does not seem to be the case that these measures typically fulfil the requirement of the technical notion of intervention that is being employed in the causal modelling literature (e.g. Woodward (2003)). This technical notion requires, among other things, that if I is an intervention on X with regard to Y, then there should not be an influence of I on Y that is not mediated by X. However, putting a criminal into prison, for example, is not only a measure that is supposed to change the criminal's motivation (reformation). It also affects the agent's actions X directly because during the time that he is in prison he is (to a certain degree) unable to act criminally (independently of his motivations to do so).

But what exactly do the variables M and X represent? We blame the criminal for performing a particular token act that is presumed to be caused by a particular token

[6] Further moral influence accounts or forward-looking accounts of moral responsibility have been provided by J. J. C. Smart (1961) and Daniel Dennett (1984). A recent defence has been provided by Vargas (2008).

motivation. And importantly we do so *after* the intentions have formed and the criminal act is performed.[7] However, Schlick describes punishment as a measure to prevent the wrongdoer (and others) from performing future crimes. Thus, M and X need to represent suitable generalizations of the token motivation and the token act that evokes the punishment. We will get back to the issue of suitable generalizations in the following section.

Retributivists, by contrast, ascribe an intrinsic value to assigning responsibility. They think that wrongdoers should be punished even if the punishment has no other positive effects. Here is a thought experiment that illustrates this intuition. Suppose there is a defendant who is convicted for rape (see Moore (1997), 100f). Suppose also that after the rape but before sentencing the defendant has lost his sexual and aggressive desires through an accident such that no danger of rape or similar crime is to be expected. Moreover, suppose that we could pretend to punish the defendant such that everyone else thinks that the defendant is in prison, but in fact he is not. Thus, the defendant is incapacitated (through the accident). Moreover, the defendant is denounced and rehabilitated through the appearance of being imprisoned, which has also the effect of general deterrence. Yet, there is an important sense in which it is not right to deal with the defendant in this way because the defendant deserves to be in prison. This is an illustration of the retributivist's *positive* desert claim.

According to the *negative* desert claim, only those may be punished who deserve it. This excludes, for example, preventive detention. Here is an example (see Moore (1997), 95). Suppose a psychiatrist finds out that one of his patients is highly likely to be very dangerous. Suppose also that the patient is being accused of crime. Suppose further that the judge is the only one who knows that the accused patient is innocent. From a consequentialist perspective the judge would have to convict

[7] Except in cases where the intention to commit a dangerous crime is known such that the harm can be prevented.

the accused in order to incapacitate him, yet, this clearly goes against retributivist intuitions because the innocent do not deserve to be punished.

There are a number of challenges that the retributivist faces. First, a retributivist, of course, needs to specify what desert consists in. Second, the retributivist needs to specify a rule for the proportionality of punishment. One prominent form of retributivism is the Biblical *Lex talionis*—"an eye for an eye, a tooth for a tooth" (*Exodus* 21: 23-25; *Leviticus* 24:17-20). This idea of payback as response in kind is of course highly controversial. However, it should be clear that this is not representative of retributivism in general. While retributivists argue hat punishment should be proportional to the wrong, they are not committed to this particular measure of proportionality (Moore, 1997, 88). Finally, another crucial question for the retributivist is how desert is to be justified. The positive and the negative desert claim surely relate to important intuitions. But the mere fact that we have these intuitions may not be sufficient to justify a moral principle. I will not try to resolve these issues here.

The important point is that while the consequentialist is concerned with the future, the retributivist is concerned with the past. In the following we will see that this has ramifications for the kinds of causal reasoning that are relevant for these theories.

5.4. Token Causal Claims and Retrospective Evaluation

One of the competitors of the interventionist account is Alicke et al.'s Culpable Control Model (CCM). According to this account, the causal judgements with regard to the pen vignette are explained with a desire to blame Professor Smith. A central difference between the CCM and the interventionist account is the following. The CCM predicts a dependence of the judgement on whether the *outcome* is negative (otherwise there is nothing the involved agents can be blamed for). By contrast, the interventionist account predicts a dependence on whether the *behaviour itself* is a norm violation or not (independent of the outcome).

Hitchcock and Knobe discuss a variation of the pen vignette in order to show the short-comings of the CCM. The new scenario is called the 'drug vignette.' It is structurally equivalent to the pen vignette but features a situation with a positive outcome:

> "An intern is taking care of a patient in a hospital. The intern notices that the patient is having some kidney problems. Recently, the intern read a series of studies about a new drug that can alleviate problems like this one, and he decides to administer the drug in this case.
>
> Before the intern can administer the drug, he needs to get the signature of the pharmacist (to confirm that the hospital has enough in stock) and the signature of the attending doctor (to confirm that the drug is appropriate for this patient). So he sends off requests to both the pharmacist and the attending doctor.
>
> The pharmacist receives the request, checks to see that they have enough in stock, and immediately signs off.
>
> The attending doctor receives the request at the same time and immediately realizes that there are strong reasons to refuse. Although some studies show that the drug can help people with kidney problems, there are also a number of studies showing that the drug can have very dangerous side effects. For this reason, the hospital has a policy forbidding the use of this drug for kidney problems. Despite this policy, the doctor decides to sign off.
>
> Since both signatures were received, the patient is administered the

> drug. As it happens, the patient immediately recovers, and the drug has no adverse effects" (Hitchcock and Knobe, 2009, 603f).

Test subjects expressed higher agreement with the claim that the doctor's signing off caused the patient's recovery than with the claim that the pharmacist's signing off caused it. Hitchcock and Knobe infer that we identify norm-violating behaviour as actual cause even if the norm violation leads to a positive outcome and, thus, no judgements of blameworthiness are involved. This effect cannot be explained by the CCM and, thus, is taken to support the interventionist account.[8]

But let us have a closer look at how exactly the drug case is explained by the interventionist. In analogy to the pen case, the drug case involves two factors that are both necessary for the outcome (with the difference that the outcome is positive). As in the pen case there is one norm-violating factor (the doctor's signing off) and one norm-conforming factor (the pharmacist's signing off).[9] In analogy to the pen case we should expect that the most suitable corrective intervention is one that targets the norm-violating factor, that is, an intervention that reminds the doctor of the hospital's drug policy.

With regard to future kidney patients such a corrective intervention on the doctor's behaviour seems reasonable. After all, the hospital's policy is backed by scientific evidence about the drug's potentially dangerous side effects. And the fact that one patient was lucky enough not to be affected by these side effects should not be a sufficient reason to change the policy. However, with regard to the retrospective evaluation of the token situation this recommendation does not seem to apply. In fact, in retrospective one should not have intervened at all because an

[8] In reply to the study by Hitchcock and Knobe, Alicke et al. 2011 give evidence to the effect that the evaluation of the outcome does play a role in the evaluation of an agent's causal role. However, the effect of the outcome being good or bad is relatively small as compared to the effect of norm-violating behaviour (only about half of the effect size). Moreover, it was not replicated in the XPhi Replicability Project (Hannikainen and Cona, 2017). By contrast, Hitchcock and Knobe's results were replicated (Phillips, 2017).

[9] One could construe the actions of the intern as another norm-conforming factor, but for the sake of the analysis we can ignore these.

intervention on either one of the involved agents would have prevented the positive outcome. I conclude that the fact that we identify the norm-violating factor is *not* to be explained by the fact that the factor is an appropriate target for intervention—at least in the token situation.

Interestingly, Hitchcock and Knobe suggest that the "hospital administrators *encourage* the attending physicians to sign off on requests to use the drug" (608, emphasis added). But this seems to be an implausible suggestion for several reasons. If it is to be understood as an hypothetical intervention on the past token situation, then it does not seem to be necessary, because in this situation the doctor signed off the request anyway. If this is to be understood as an intervention on the doctor's future behaviour, then it is problematic because it puts future patients at risk of suffering dangerous side effects. Hitchcock and Knobe's suggestion is also in conflict with their theory which says that norm-violating factors are suitable targets for *corrective* interventions. Encouraging the doctor's behaviour instead would amount to stabilizing norm-violating behaviour.[10]

Thus, there is a mismatch between what is to be done in future instances of this kind of situation and the evaluation of the token situation. And this mismatch seems to raise difficulties for the interventionist view. By contrast, such problems do not need to arise for interpretations of the case in terms of responsibility. There is not necessarily a conflict between stating that the doctor is responsible for the positive outcome and stating that, in future instances, the doctor should not violate the hospital's policy.

Why is the responsibility view able to account for this case while the interventionist account faces problems? The key difference between the two approaches is that (retributive) responsibility involves the *retrospective* evaluation of the outcome of a

[10] Alternatively, one could understand Hitchcock and Knobe's idea of encouraging the doctor's behaviour as a suggestion to change the underlying norm. But then, again, it does not seem to be recommendable to change the policy because of a single instance where the side effects did not occur.

situation. The interventionist account, by contrast, is concerned with the *prospective* evaluation of suitable strategies. The interventionist justification for emphasizing norm-violating factors is derived from the fact that these factors are suitable targets for *future* intervention. In the pen case, for example, Hitchcock and Knobe do not seem to suggest that we literally intervene on Professor Smith's past behaviour in order to make the problem of pen shortage undone. Instead, the token situation is interpreted in the light of possible interventions that would prevent future instances of pen shortage.

This seems to be related to the fact that we cannot literally change the past. But couldn't we interpret the interventionist position as suggesting *hypothetical* interventions? Hypothetical interventions are interventions that one should have applied or interventions that one should apply if one could travel back in time and literally make the problem undone. But even then the interventionist evaluation seems to have prospective character. Given the knowledge of the drug's dangerous side effects (but not the *post hoc* knowledge of the lucky recovery) one should have intervened on the doctor such that she would not have signed off the request for the drug.

By contrast, the retrospective evaluation is sensitive not only to the action of the doctor but also the actual result of the action. Normally, the doctor's violating the norm is blameworthy because it puts the patient at risk. However, in this specific case the doctor's blameworthiness is mitigated because the adverse side effects did not occur. In fact, the patient may even have an inclination to praise the doctor because she cured her. Thus, the evaluation of the doctor's behaviour depends on aspects of the situation that were neither foreseen by the doctor nor under the doctor's control. This phenomenon has been described as moral luck (see e.g. Nagel (1979, Williams (1981). Moral luck is widely believed to have an impact on our retrospective evaluation of an agent's responsibility. This is reflected, for example, by the fact that murder is commonly assessed differently than attempted murder that failed due to external reasons.[11] This kind of consideration does not seem to have an analogue in the prospective evaluation of possible ways to bring about a certain effect.

5.5. Late Preemption

In this section we shall turn to claims of actual causation in the context of redundancy. What is the use of these claims from the interventionist perspective? Clearly they do not entail simple recommendations for the best target for intervention as in the pen vignette and structurally similar cases. The reason is that intervening on the actual cause in cases with redundancy will not prevent an undesired outcome. The outcome will be sustained by the alternative process(es).

However, in Chapter 4 we have seen that in such cases claims of actual causation provide information that is useful if the intervening agent can apply multiple interventions. If a factor is a path-changing actual cause, for example, its effect can be prevented if we intervene on the path-changing actual cause and combine that intervention with a secondary intervention that counteracts the adverse consequences of the primary intervention.

In this context a particular challenge has arisen from cases of late preemption (see the discussion in Chapter 4). On the one hand, there is an intuitive difference between the actual cause (the preempting factor) and the alternative causal processes (that correspond to the preempted factor(s)). On the other hand, it is not clear what the practical value of this intuition is from the interventionist perspective. Consider the Suzy-Billy case. Suzy throws a stone at a bottle and destroys it. Just after Suzy's throw (and before the bottle is destroyed) Billy throws his stone. If Suzy had not

[11] It is a different question whether we are justified to assume that there is moral luck. Maybe we shouldn't assess murder differently than attempted murder that failed due to external reasons.

hit the bottle, then the bottle would still have been intact upon the arrival of Billy's stone and Billy's stone would have destroyed it. There is a clear intuition that Suzy's stone is the actual cause of the bottle's shattering but not Billy's stone. But what does this imply for an agent who wants to save the bottle?

In order to save the bottle, the agent has to intervene, first, on Suzy's stone such that it does not destroy the bottle. One consequence of this intervention is that the bottle is still intact upon the arrival of Billy's stone. So the agent needs to apply a second intervention such that Billy does not hit the bottle. However, it is not clear why the agent would need to know that in the original situation Suzy's rather than Billy's stone would destroy the bottle in order to come up with this strategy. The agent has to intervene on both Suzy and Billy in the same way, and this seems to be independent of whether Suzy or Billy would be the actual cause of the bottle's shattering.

Do the minor differences in timing play a role? Presumably we have $|t_{SH} - t_{BT}| < |t_{BH} - t_{ST}|$. This would leave more time between the two interventions if we were to intervene first on Suzy and then on Billy. But why should such minor differences in timing be relevant? Moreover, the differences in timing are contingent upon the stones' velocities. If Suzy's stone has a higher velocity it can hit the bottle even if Suzy throws later than Billy. Thus, the differences in timing do not support an interventionist interpretation.

Does the same problem arise in cases of early preemption? That is, could one reverse one's strategy in the case of early preemption as well? Surely one could save the victim by first intervening on the supervisor and then intervening on the trainee just before he pulls the trigger. But this requires the intervening agent to know in advance that the supervisor would step in as a result of an intervention on the trainee. In the foregoing chapter I have argued that cases of early preemption are often so difficult to handle because this kind of counterfactual information may not be available. In cases of late preemption this kind of counterfactual knowledge tends to be more accessible since there are two processes that represent independent threats to the goal.

I conclude that from the interventionist perspective it is difficult to explain the practical value of drawing a distinction between preempted and preempting factors in examples of late preemption. Of course, the claim of actual causation provides *some* information. In the Suzy-Billy case it tells us whose stone destroyed the bottle. The value of this additional information, however, may be explained from a perspective other than that of the interventionist. In particular, it seems plausible to say that the value of this information is to be explained by its relevance for assigning responsibility. The claim that Suzy's throwing her stone is an actual cause of the bottle's shattering justifies ascribing responsibility for destroying the bottle to Suzy. By contrast, Billy can only be held responsible for intending to hit the bottle.

Is the difference between Suzy's hitting the bottle and Billy's attempt to hit the bottle relevant for our actual practices of blaming and praising? Sometimes it is relevant. Suppose, for example, that Suzy and Billy participate in a bottle shattering competition (as described by Hall (2004, 235)). In this context we do care about Suzy's being the actual cause of the bottle's shattering because we want to give her credit for her achievement. Sometimes it is not relevant. Suppose, for example, Suzy and Billy are members of a firing squad. Suzy shoots a little earlier and kills the victim immediately. Billy shoots a little later but before Suzy's bullet reaches the victim. Suzy and Billy are most likely seen to be equally blameworthy for the victim's death. In fact, one of the purposes of execution by firing squad is to create a sense of diffusion of responsibility among its members. But in this kind of case the corresponding actual causal claim is not only irrelevant for matters of responsibility—it is irrelevant *tout court*. So, I claim that *if* we care about actual causation in late preemption scenarios, then there are cases where the interest is better explained in terms of responsibility than in terms of intervention.

Moreover, whether actual causation in late preemption is related to questions of responsibility depends again on one's take on moral luck. Consider the following case. Suppose Suzy and Billy are assassins that both attack an innocent victim. They have been sent on their missions by different clients and they attack without knowing of each other. Suzy and Billy pull the triggers of their guns almost at the same time but Suzy shoots a little earlier. Billy pulls the trigger while Suzy's bullet is propagating towards the victim such that Billy's bullet does not arrive before the victim is killed through Suzy's bullet.

Does it matter that Suzy murdered the victim while Billy only attempted to murder the victim? If one rejects the idea of moral luck, then our evaluation of Billy's actions should not depend upon whether he is an actual cause or not. The only reason for Billy's not being an actual cause is that Suzy killed the victim a split second earlier. And this is an aspect of the situation that was neither under the control nor foreseen by Billy. Correspondingly, from the prospective perspective of possible interventions the difference should not matter. In order to save future victims that are being attacked by Billy and Suzy we need to discourage the behaviour of both Billy and Suzy. However, for the retrospective evaluation of the assassins' actions moral luck does seem to play a role. And this corresponds to the fact that the difference between preempting and preempted factor is considered to be important.

5.6. Conclusion

In this chapter I have examined the scope of interventionist accounts of the function of actual causation. It is a commonly accepted assumption among interventionists that a key role of actual causation is the *post hoc* evaluation of responsibility. Thus, the prospects of interventionist approaches to provide an exhaustive account of the function of actual causation depend on whether intervention and responsibility align or not. In order to assess the relation between intervention and responsibility I have introduced a distinction between moral influence and retributivist accounts of responsibility. There is a close analogy between moral influence accounts and interventionist accounts. In both cases the evaluation of causes has prospective character, even if the causes occurred in the past. Retributivist responsibility, by contrast, involves a retrospective evaluation of past causes.

This has ramifications for the scope of interventionist accounts of the function of actual causation. I have discussed two instances where interventionist accounts face limitations and where a better explanation is provided from the perspective of responsibility. First, in the situation described by the drug vignette we identify the doctor's norm-violating behaviour as actual cause. This is problematic because a corrective intervention on the doctor's behaviour would have had adverse consequences for the outcome of the situation. Second, in cases involving late preemption it is difficult to motivate the distinction between preempting factors and preempted factors from an interventionist perspective. In order to prevent the outcome an agent would have to intervene on both kinds of factors in the same way.

Part III.

Consequences

6. Actual Causation in the Law

In the law the notion of actual causation goes back to the American Legal Realists of the early 20th century. The Legal Realists were concerned with delineating the factual and principled elements of legal inquiry from those that depend upon context-sensitive considerations involving norms and policy. They introduced the notion of "actual causation" and contrasted it with the notion of "proximate cause" which refers to the context-sensitive and norm-related elements of legal inquiry. In his first treatment of the concept Judea Pearl refers to this tradition in the law, arguing that in the law actual causation is taken to be "the ultimate criterion" (2000, 309) for responsibility. However, in the preceding chapters we have seen that, ironically, contributors to the causal models literature (including Pearl) have used the term actual causation in order to describe context-sensitive and norm-dependent modes of causal reasoning.

In this chapter my aim is to situate the results of the foregoing chapters within the legal debate on actual causation. More specifically, I will disentangle the apparently conflicting takes on actual causation by employing a disambiguation between two kinds of context-sensitivity introduced in Chapter 4. Context-sensitivity$_1$ concerns our willingness to consider individual variables' values as default states or deviant states. Context-sensitivity$_1$ features in selecting causes as salient factors from a set of jointly sufficient background factors. This kind of selection can be described as depending on considerations of normality: we typically identify those factors as salient that are in some sense abnormal. An example for this kind of context-sensitivity is the following. In most cases the short-circuit is identified as the cause of the fire while the presence of oxygen is a mere background condition. However, there are also contexts where the presence of oxygen is identified as actual cause and the short-circuit is a background condition, for example, if the fire occurs in a chamber that is supposed to be evacuated.

Context-sensitivity$_2$ concerns our willingness to consider complex counterfactuals that can be described as involving combinations of interventions. An example has been given in Chapters 3 and 4. In most circumstances we are willing to take seriously the possibility that (in violation of the structural equations) the supervising assassin does not kill the victim even though the assassin in training failed to pull the trigger of his gun. Therefore, we consider the assassin in training to be an actual cause of the victim's death. However, there may also be circumstances where a failure of the supervising assassin is only a far-fetched possibility. In such circumstances the assassin in training would not be identified as actual cause.

After a brief review of the historical debate on actual causation in the law I will address Richard Wright's version of the NESS account as a contemporary instance of the realist camp. The NESS account states that a cause is a necessary element of a sufficient set of conditions for the effect. It was first developed by Hart and Honoré (1959) and John L. Mackie's (1965) INUS condition is a development of it that is better known among philosophers.[1] Wright embeds the NESS account as a criterion of actual causation in a three-phase scheme of legal inquiry. I will argue that while Wright may thereby have succeeded in eliminating context-sensitivity$_1$, he encounters problems with context-sensitivity$_2$. In my argument I will focus on cases of preemptive prevention. These are cases that concern factors that prevent a

[1] According to Mackie, a cause is an INUS condition, that is, an insufficient but necessary element of a set that of factors that is unnnecessary but sufficient for the effect. Honoré declares that Mackie "applied our idea" (Honoré, 1997, 365).

particular outcome that otherwise would have been prevented by a second factor. Wright has employed cases of preemptive prevention in order to clarify cases that involve, for example, the nonuse or misuse of defective or missing safety devices. If a safety device is defective, then it is prevented from operation. But if the defective device is not even put to use, this means that it is prevented from operation anyway. We say that the prevention through nonuse preempts the prevention through the device's being defective.[2] Causal inquiry has been considered particularly difficult in these kinds of cases because standard tests such as the but-for criterion fail. Wright offers a principled (context-independent) account for such cases. I will argue that the account achieves such a treatment of preemptive preemption only at the price of committing to a class of highly implausible causal claims. This suggests that context-sensitivity$_2$ cannot be eliminated from causal inquiry in the legal context.

The chapter is structured as follows. In section 6.1 I will provide a brief review of the debate on actual causation in the law in the beginning of the 20th century. This section will motivate the Legal Realist's distinction between actual and proximate causation. In section 6.2 I will present first doubts against a principled account of actual causation. These doubts have been put forward by Wex Malone and by Hart and Honoré. The discussion of these two accounts will motivate a closer look at Wright's account, a contemporary realist account that opposes the arguments put forward by Malone and Hart and Honoré. In section 6.3 we will see that Wright distinguishes three phases of legal inquiry: (1) tortious-conduct inquiry, (2) application of the actual-cause requirement, and (3) application of the proximate-cause criterion. According to Wright, the first and the third phase are affected by context-sensitive and norm-dependent considerations. The second stage which,

[2] We will see that this kind of case also involves double prevention: the defective device is prevented from preventing some harm. In this sense these legal cases differ from cases of preemptive prevention that have been discussed by McDermott (1995) and Collins (2004).

according to Wright, concerns causal inquiry in the proper sense is not so affected. I will highlight how this account manages to sideline issues related to context-sensitivity$_1$. In section 6.4 we will have a closer look at Wright's NESS criterion for actual causation which promises to provide a principled account of actual causation. In section 6.5 I will introduce cases of preemptive prevention. Finally, in section 6.6, I will show that these cases raise difficulties for Wright's account because they involve context-sensitivity$_2$. The difficulties indicate that context-sensitivity$_2$ remains an essential feature of actual causation in the legal context, even in Wright's three-phase framework.

6.1. Actual and Proximate Cause

Theories of causation in the law have mainly focused on the law of tort, a branch of private law.[3] Tort law regulates cases that are prosecuted by the victim of some wrong in order to seek (often financial) compensation from the person who hurt the victim. Thus, tort law differs from criminal law that regulates cases prosecuted by the state and that may result in a sentence of punishment. The reason for the focus on tort law is that "a large part of tort law consists of but one injunction: do not unreasonably act so as to cause harm to another" (Moore, 2009, 83). The meaning of this injunction, of course, depends heavily on the meaning of "to cause." In this regard tort law differs from criminal law that specifies in much more detail which actions are prohibited or required.

A common test for causation in the law is the but-for test, also referred to as a test that determines the *conditio sine qua non*. This test essentially construes causation along the lines of counterfactual dependence: would the victim's injury have been avoided if the defendant's actions had been different? But of course the but-for

[3] I focus here on a debate that is mostly concerned with causation in the American legal system. In the German legal system the relevant branch is called "Deliktsrecht."

test is an incomplete method for determining the defendant's liability. For we are typically held liable not for all adverse consequences that depend counterfactually upon our actions. Here is an example:

> "Suppose a parent, D, fails to control their two-year-old infant who runs out into the path of a moving vehicle which swerves and breaks the leg of a pedestrian, P. On the way to hospital, the ambulance carrying P is struck by lightning and P is seriously burnt" (Stapleton, 2008, 448).

The pedestrian's broken leg and burns both depend counterfactually on the parent's negligent behaviour. Yet the parent will be held liable for the broken leg but not for the burns since the burns are too remote from the parent's conduct. The example illustrates that liability requires that the defendant's actions stand in a relation to the victim's injury that is more demanding than the relation probed by the but-for test. This requirement is commonly thought to be captured by the notion of *proximate cause*.

Discussions of the meaning of proximate cause go back as far as Francis Bacon's *Maxims of the Law* and saw a first climax in the early 20th century (see e.g. Smith (1912); Beale (1920); Carpenter (1932)). Contributors to this debate aimed to find "definite principles of law by which the determination of proximity is to be regulated" (Beale, 1920, 636). An example for such a criterion is Smith's substantial factor criterion. According to this criterion, the "*effect* of defendant's tort must have appreciably continued; either down to the very moment of damage; or, at least, down to the setting in motion of the final injurous force which immediately produced (or preceded) the damage" (1912, 310f). For example, one may think that the effect of the parent's negligence is the child's running into the path of the moving vehicle and thus affects the motion of the vehicle which is the final injurous force in producing the pedestrian's broken leg. By contrast, there is no such continuation of the defendant's tort to the burns.

However, there were doubts that a principled account of proximate cause could be given. These doubts were voiced prominently by the American Legal Realists (e.g. Edgarton (1924); Green (1929)). Green, for example, argues with regard to Smith's substantial factor criterion that "the answer [to whether the defendant's conduct was a substantial factor] is only to be had in the judgment of the particular tribunal [...] to which the problem is allocated [...]" (1929, 604). Another instance of this criticism concerned the foreseeability criterion (according to which, for example, the broken leg is a proximate cause because it is foreseeable, but not the burns). According to Green, what exactly is foreseeable is not to be determined on the basis of principles. Instead it depends upon stipulations about what an "ordinary prudent person" can foresee and this, in turn, "must be defined and oriented by the "circumstances of the particular case [...]." Green concludes that both the criterion's "vice and virtue lie in the fact that it may count for anything or for nothing. Its function is similar to that of a joker in the game of poker" (1929, 612).

The Legal Realists suggest a clear separation of two elements of a legal inquiry. First, there is the factual part of the inquiry that is considered to be truly causal. This element of the inquiry is thought to establish whether the defendant's conduct stands at all in a causal relation with the victim's injury. If such a relation exists, then the defendant's conduct is considered to be an *actual cause* (also: "material cause", "cause-in-fact", or "but-for cause") of the injury. Second, there is the proximate cause inquiry. The Legal Realists are sceptic with regard to any principled approach to this relation and argue that it is a matter of policy.

The Legal Realists also argue that the distinction between actual cause and proximate cause has a profound influence on the actual procedure in court. According to Green, "[t]here is extremely little work [...] to do" in the first part of the inquiry "for normally causal relation is so clear that a judge would not be warranted for submitting it to a jury" (1929, 607). However, Green further argues that since "judges do not recognize what a narrow problem causal relation is [...] in almost every case submit some other problem which should not be submitted to the jury at all" (ibid.). So the distinction, according to Green, is so important because it affects which issues are given to the jury to decide.

6.2. Challenging Principled Accounts of Actual Causation

An interesting twist to the Legal Realist's account is introduced by Wex Malone. In an influential article from 1956 Malone argues that policy-related issues do not only affect proximate causation but also judgements of actual cause (which is called "simple cause" or "cause-in-fact" in Malone's terminology):

> "I find that even within reference to this issue of simple cause the mysterious relationship between policy and fact is likely to be in the foreground. [...] [I]t will be demonstrated that policy may often be a factor when the issue of cause-in-fact is presented sharply for decision, much as it is when questions of proximate cause are before the court" (61).

Malone's main argument for the effect of policy on matters of actual causation concerns cases like the following. Suppose an elderly worker with a heart ailment happens to die from a heart attack while performing some trivial task for his employer. A doctor, Malone argues, will not view the trivial task as a cause of the worker's death. The doctor "cannot escape forming associations between events that will comport with the purposes of his profession" (63). These purposes concern diagnosing, curing, or otherwise preventing the adverse effects of diseases. Therefore, the doctor "will likely envision as causes only those factors with which he can deal in diagnosing, in curing or in seeking to forestall future occurrences of this kind for other persons" (ibid.). By contrast, a judge who is concerned with the workers' compensation statute may well think of the trivial task as a cause of the worker's death. For "[h]e likely will be impressed by the law's desire to throw its protection around the susceptible and aged worker as well as the one who is in sound health" (64).

Malone takes this example to show that

> "[m]uch misunderstanding between lawyers and physicians could be obviated if members of both professions would realize that "simple" causation is not merely an abstract issue of fact and that the resolution of the cause problem depends largely upon the purpose for which cause is to be used. What is *a* cause for the judge need not be *a* cause for the physician. It is through the process of selecting what is to be regarded as a cause for the purpose of resolving a legal dispute that considerations of policy exert their influence in deciding an issue of cause-in-fact" (64).

Here Malone describes an instance of what I have earlier labelled as context-sensitivity$_1$: selecting a salient cause from a range of background conditions. The problem of causal selection in legal inquiry and beyond was treated in a more systematic way by Hart and Honoré (1959). In Chapter 2 we have seen that, according to Hart and Honoré, the two main criteria for whether a but-for condition is identified as a cause or merely as a background condition are, first, that the condition be abnormal or, second, be a voluntary action. What is normal, according to Hart and Honoré, depends upon the specific effect that is being considered and it also depends on the pragmatic interests of the causal reasoner. We have also seen that a weaker version of Hart and Honoré's normality criterion has had major influence on discussions on actual causation in the causal modelling literature. Thus, it seems like the distinction between actual causation as a purely factual relation and proximate causation as a norm-dependent relation cannot be drawn: both concepts involve context-sensitive considerations about normality. But is it legitimate to consider normality as part of the notion of actual causation in the law?

6.3. Three Stages of Legal Inquiry

Wright advocates a strict separation between the causal and the non-causal elements of a legal inquiry and, thus, stands in the tradition of the Legal Realists. He acknowledges the insights by Malone and Hart and Honoré as highlighting relevant elements of the legal inquiry. However, he argues, the norm-related elements are not part of the *causal* part of the legal inquiry. More specifically, Wright distinguishes three stages of the legal inquiry as follows.

(1) The first stage is the tortious-conduct inquiry. The aim of this part of the inquiry is to identify tortious aspects of the defendant's conduct that potentially caused the injury. Such tortious aspects and not the defendant's overall conduct or other factors will be analysed in the subsequent stages of the inquiry. This stage of the analysis is norm- and policy-dependent because whether a potential cause is tortious or not is a matter of norm and policy.

(2) The second stage is the application of the actual-causation requirement. The aim of this stage is to determine which of the tortious aspects actually caused the injury. According to Wright, this is the only part of the inquiry that deserves to be called 'causal.' Wright's criterion for actual causation is the NESS test, which will be discussed shortly.

(3) The third stage is the proximate-cause inquiry. This stage determines whether liability is reduced or eliminated because of contributing factors other than the defendant's tortious conduct (as in the case of the burns as a result of the lightning that hit the ambulance). This third step, again, is a matter of policy and, according to Wright, should not be understood as being concerned with causation at all but merely with liability.

How does this scheme help to separate the context-sensitive and norm-dependent aspects of legal inquiry from the causal aspects that supposedly can be treated in a principled way? First, the distinction between the actual-cause requirement and the proximate-cause inquiry reflects the distinction made by the earlier legal realists. Second, there is a tortious-conduct inquiry that precedes the application of the actual-cause requirement. What counts in a legal inquiry, according to Wright, is not whether the overall conduct of the defendant (operating a hotel, driving a car) stands in a causal relation to the harm. What counts is that the tortious aspects of it (failure to provide a fire escape, excess speed) stand in a causal relation to the harm. Thus, what is submitted to the NESS test is behaviour that violates certain norms (such as the duty to provide a fire escape when operating a hotel or conforming to the posted speed limit while driving). The distinction between the tortious-conduct inquiry and application of the actual-causation requirement can thus be understood as a response to the objections put forward by Malone and Hart and Honoré. The context-sensitive$_1$ questions of selection that Malone and Hart and Honoré are concerned with are taken into account by Wright, but he describes them as not being part of the causal part of the legal inquiry.

One might worry that selecting tortious conduct is itself a procedure that rests upon certain causal assumptions. The legal inquiry is not concerned with all kinds of tortious conduct but with tortious conduct that is plausibly related to the harm. A clean separation may thus not be as straightforward as suggested by the three-phase scheme. Yet, in the following we shall grant that Wright's account achieves a separation of considerations associated with context-sensitivity$_1$. We shall now turn to issues related with context-sensitivity$_2$.

6.4. The NESS Criterion

In order to assess the role of context-sensitivity$_2$ let us first have a closer look at the NESS criterion, which Wright takes to be the basis of the second stage of the inquiry. Wright argues that

> "[t]he essence of the concept of causation [...] is that *a particular condition was a cause of (condition contributing to) a specific consequence if and only if it was a necessary element of a set of antecedent actual conditions that was sufficient for the occurrence of the consequence*" (1790, emphasis original).

The acronym NESS summarizes what is considered to be the core of the account, namely, that a cause is a "*n*ecessary *e*lement of a *s*ufficient *s*et" (1790) of conditions for the effect. The NESS account goes back to Hart and Honoré (1959) and in the philosophical literature its most prominent development was provided by Mackie in the form of the INUS condition.

I shall illustrate the NESS criterion by showing how it deals with the well-known cases of redundant causation. We shall begin with symmetrical overdetermination. Here two events c_1 and c_2 both cause the effect e. The NESS criterion straightforwardly accounts for this kind of case. The causes c_1 and c_2 are simply taken to be necessary elements of two separate sets that are individually sufficient for the effect. For example, if two lightning bolts cause a forest fire, each of the lightning bolts is a necessary element of a set that is sufficient for the occurrence of the fire—a set comprising, e.g., the presence of inflammable material, oxygen . . . at the place where the lightning bolt strikes).

Next we turn to preemption. In preemption cases a cause event c_1 preempts an alternative event c_2 that could have caused effect e in the absence of c_1. The problem with this kind of case is that both c_1 and c_2 are necessary elements of a sufficient set for e. For example, in Backup (see Chapter 2) the trainee's pulling the trigger of the gun is a necessary element of a set that is sufficient for the victim's death (including the fact that the victim is standing at the right spot to be hit by the bullet,...). Analogously, the supervisor's shooting is a necessary element of a sufficient set for the victim's death. According to Wright, however, the supervisor's shooting does not count as a cause in the NESS account because it is not "a part of any set of *actual* antecedent conditions that was sufficient for [the victim's death]" (Wright, 1985, 1795). That is, the supervisor's pulling the trigger is not a cause because in the actual situation the supervisor simply did not pull the trigger.

But note a potential complication: while the supervisor does not shoot, earlier events in the causal chain do occur. Consider, for example, the supervisor's determination to shoot if the trainee doesn't. This will turn out to be a NESS condition for the victim's death (in a set that leaves out the trainee's shot).

Wright proposes a similar treatment of late preemption. In late preemption both c_1 and c_2 do occur, but only c_1 causes e because c_2 occurs too late. Suppose both the supervisor and the trainee shoot but only the trainee hits and kills the victim because he pulled the trigger a little earlier. Thus, the supervisor did pull the trigger in the actual situation. Yet, according to Wright, the supervisor's action is still not part of a set of actual antecedent conditions because other necessary elements of the set are not present in the actual situation. One requirement is that the victim is alive at the time when she is hit by the supervisor's bullet. But the victim is already dead because the trainee's shot killed her immediately (by stipulation).

But note that this analysis trades on an ambiguity. The victim's being alive when the supervisor's bullet hits is necessary for the supervisor's bullet to *cause* the victim's death. But the NESS account cannot appeal to this, on pain of circularity. The question is whether the supervisor's shot is a NESS condition for the victim's death. And this appears to be true. The victim being alive when the supervisor's bullet hits is *not* necessary for the victim's death.

Thus, Wright's NESS account appears to face problems if applied to the standard test cases. But these problems shall not concerns us in the following. Instead we shall focus on instances where the NESS account faces difficulties arising from context-sensitivity$_2$.

6.5. Preemptive Prevention

Consider the following case of preemptive prevention (McDermott, 1995). Suppose a fielder catches a cricket ball that was flying in the direction of a window. Between the first fielder and the window there is a second fielder. The second fielder would have caught the ball if the first fielder had not caught it. Did the first fielder's catching the ball prevent the window's shattering? Intuitions are ambiguous. In a way it did not: the ball would not have hit the window irrespective of the first fielder's action. In another way it did: if neither fielder had been present the window would have been shattered. But which one of the two fielders did the preventing? Clearly the first fielder because the second fielder did not contribute.

Consider the following causal model (figure 6.1). The first fielder catches the ball ($FC = 1$) if it is thrown ($TB = 1$). As a result the ball cannot be caught by the second fielder (represented by $Z = TB \wedge \neg FC = 0$) and the window does not break ($BW = TB \wedge \neg FC \wedge \neg Z = 0$).

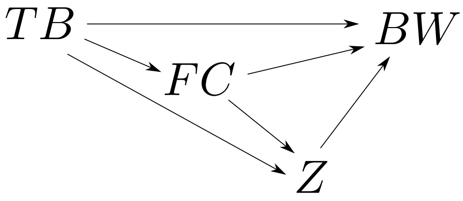

Figure 6.1.: Preemptive prevention.

Causal model accounts of actual causation such as the one proposed by Hitchcock (2001) nicely reflect the ambiguity of this case. The existence of an active causal route between FC and BW depends on whether we are willing to take seriously scenarios where the intermediate variable Z is held fixed at its actual value *given that FC* is intervened upon—which is why this is an instance of context-sensitivity$_2$. Given that the fielder does not catch the ball how likely is it that the second fielder does not catch the ball either? Presumably, the second fielder is fallible. Thus, we should not disallow the combination $FC = 0 \wedge Z = 0$. Consequently, $\langle FC, BW \rangle$ is an active route and the first fielder's action is an actual cause of the window's being saved. In the following we shall refer to this scenario as the fielder-fielder scenario.

The ambiguity is even clearer if we consider a variation of the case suggested by Collins (2004). Here the second fielder is replaced by a solid brick wall (henceforth we shall refer to this scenario as the fielder-wall scenario). In this version of the story variable Z represents whether the wall blocks the ball or not. Again we ask the question: given that the first fielder does not catch the ball, how likely is it that the wall does not block the ball? This is much less likely than a failure of the second fielder. Again we should allow the combination of $FC = 0 \wedge Z = 0$ only to the degree that we think the wall could fail to block the ball given that the fielder does not catch it. And since this scenario is extremely unlikely, we should disallow it. Correspondingly, it is much less plausible to say that there is an active route linking the first fielder's actions and the window's being saved. This explains why we are much more reluctant to consider the fielder to be an actual cause in this scenario.

How would we model the fielder-wall scenario such that the fielder is not identified as an actual cause? Note that in our first model of the case Z represents the fact whether the wall blocks the ball or not. There are two possible ways in which the non-blocking would occur. Either (1) it occurs as a result of the fielder's catching the ball, which is a reasonable scenario or (2) it occurs even though the fielder fails to catch the ball. This is the implausible scenario. Let us adjust the set of variables $\mathcal{V}$ such that this implausible scenario is excluded. The easiest way to do this is to introduce a variable W that represents the fact that there is a wall if it takes on value 1 (if there is no wall it has value 0). Whether there is a wall is independent of whether the fielder catches the ball.[4] Moreover, the window's breaking depends on the fielder's catching the ball and the presence of the wall as

[4] This is the essential difference to variable Z which represents whether the wall blocks the ball, that is, has physical contact with the ball, which *does* depend on the fielder's actions.

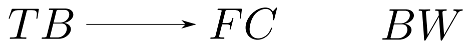

Figure 6.2.: An alternative representation of the fielder-wall scenario.

follows: $BW = TB \wedge \neg FC \wedge \neg W$. The case states that there is a wall and it seems like an extremely far-fetched scenario that this wall could suddenly disappear. So, it seems like the possibility that $W = 0$ should not be part of the causal model, which means that we should set $W = 1$ and background the variable.[5] But this means that $BW = 0$ no matter whether the ball is thrown and no matter what the fielder does. This is reflected by the model in figure 6.2 that simply represents BW as an independent variable.

According to this model, the fielder is not an actual cause. Neither is the wall an actual cause. In fact, according to this model, there is *no* actual cause of the window's remaining intact. This appears to be plausible because the ball's being thrown and then being caught by the fielder is a causal process that is independent of the window's remaining intact. This independence is explained by the causal structure of the situation. More specifically, the independence is explained by the presence of the wall which imposes a constraint on the ball's possible trajectories. This is a causal explanation because the constraints imposed by the wall are causal constraints.

For those who consider it plausible that in the fielder-wall scenario the fielder is an actual cause Collins has a third scenario that involves an even more far-fetched scenario (2004, 112f). Suppose someone throws the ball, aiming at Halley's comet. The fielder catches the ball. If the fielder had not caught the ball it would not have collided with Halley's comet because of Earth's attractive gravitational force and Earth's atmosphere (henceforth we shall refer to this as the fielder-gravitation scenario). Our taking the fielder's catching the ball to be an actual cause depends on the degree to which we are willing to entertain the scenario that Earth's atmosphere

[5] Alternatively, we could include the variable but set $W = 1$ as the default.

and gravitational field are not present. This, of course, is pretty far fetched. Thus, it would be appropriate to describe this case with a model analogous to the one in figure 6.2, replacing *BW* with a variable that specifies whether Halley's comet is hit (and a backgrounded variable representing the presence of Earth's atmosphere and gravitational field).

6.6. NESS and Preemptive Prevention

Preemptive prevention cases have been employed in the literature on actual causation in the law in order to clarify cases that involve an injury that is caused by the theft, nonuse, or misuse of defective or missing safety devices. An example is *Saunders System Birmingham Co v Adams*:[6]

> "C negligently failed to discover and repair defective brakes in a car that he rented to D, and D negligently failed to try to use brakes to avoid running into P. It is assumed that the injury to P would have been avoided if and only if C had repaired the brakes and D had tried to use them" (Wright, 1985, 1801).[7]

In the following we shall have a closer look at Wright's treatment of the braking case. Wright has repeatedly (1985; 2001; 2011) argued that the NESS criterion identifies D's failed attempt to use the brakes as the actual cause in a clear and non-ambiguous way. According to Wright, the case is to be described as an example of overdetermined negative causation. Negative causation means that there is a causal process that is interrupted by some actual cause. Overdetermination reflects the

[6] 117 So 72 (Alabama, 1928).

[7] Further instances concern cases where some harm occurs as a result of a failure to warn. Suppose a "product manufacturer fails to put a required warning on a conspicuous product label containing other warnings. The product user fails to read the label, and harms a bystander by using the product in a way that would have been prevented had the omitted warning been provided, read, and heeded" (Fischer, 2005-2006, 300).

fact that there are several factors that could have interrupted the process. According to Wright, "when analysing overdetermined negative causation, it is critically important to focus on the sequencing of the steps in the positive causal process that failed, in order to determine at which step it failed" (2011, 317). The relevant causal process, according to Wright, is the braking process.[8] Wright describes this process has having several stages, the first stage being the driver's applying force to depress the brake pedal. As a consequence, a lever is being operated which leads to hydraulic brake fluid being transmitted into the system. Ultimately this leads to braking pads being pushed against parts of the wheel such that the wheels are slowed down through friction. Because of D's failure to brake, the causal process does not reach the later stages. This is why, according to Wright, the failure of the brake system (which had occurred at one of the later stages) is not part of a set of actualized sufficient conditions and, thus, not an actual cause.

This treatment of the braking case was criticised by Fischer (1992; 2005-2006) and Stapleton (2008) who argue that it is arbitrary to choose the force applied to the brake pedal as the start of the process that lead to plaintiff's injury. Instead one could just as well stipulate that the relevant process begins with the failure to repair the brakes. Since the brakes are not functional at the time where D should have operated them, D's failure to brake is not a necessary element of the set of conditions that lead to the accident. This, according to the critics, would mean that the defective brakes are the actual cause.

In order to counter this objection Wright draws an analogy to McDermott's and Collins's thought experiments regarding preemptive prevention. Before we address the feasibility of Wright's reply, let me make the analogy as precise as possible by providing a causal model (see figure 6.3). This will reveal that between McDermott's and Collins's cases, on the one hand, and the braking case on the other hand, there

[8] I disagree: the positive causal process is the car approaching P. The braking process would have prevented this. See the discussion below.

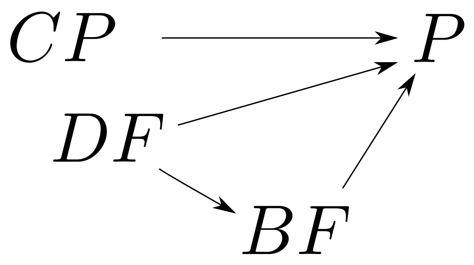

Figure 6.3.: The braking case represented in partial analogy to preemptive prevention.

is a decisive difference. $CP = 1$ shall represent the fact that the car approaches P ($CP = 0$ otherwise). $DF = 1$ shall represent the fact that D fails to depress the brake pedal ($DF = 0$ if D does not fail to depress the brake pedal). $BF = 0$ shall represent the fact that the brakes do not fail ($BF = 1$ if they do fail).[9] Finally, $P = 0$ shall represent the fact that P is injured ($P = 1$ otherwise). In order to implement the analogy invoked by Wright, we would need to provide structural equations that are isomorphic to the fielder cases.

P will be injured if the car approaches and either D fails to operate the brake or the brake fails: $\neg P = CP \wedge (DF \vee BF)$. This equation is structurally similar to the corresponding equation in the fielder-fielder case. Moreover, we may assume that the brake's failure depends on D's actions as follows: $BF = \neg DF$, meaning that the brakes fail if D operates them. This would be in analogy to the fielder-fielder case as the second fielder's actions depend on the first fielder's actions.[10] However, note also that the first differences to the fielder-fielder case arise from the fact that D's failure to operate the brake and the brake's failure are independent of the car's approaching P.

Then the analogy, presumably, is supposed to work as follows. In the fielder-fielder scenario (as well as the fielder-wall scenario) there is a positive causal pro-

[9] In the actual scenario the brakes do not fail because they are not even put to work. This is analogous to the second fielder not catching the ball because the ball did not even arrive at her position.

[10] For now, let us assume that this is an appropriate representation. Shortly, I will argue to the contrary.

cess: the ball's approaching the window. In analysing the case of overdetermined negative causation we have to follow the sequence of this positive causal process and identify the first instance where it is interrupted. The ball's approaching the window is interrupted by the first fielder (or the only fielder, in the fielder-wall scenario) and, thus, the fielder's catching the ball is the actual cause. Likewise, according to Wright, in the braking case there is a positive causal process (presumably, the braking) which is interrupted by the failure to operate the brakes, which makes D the actual cause of P's being injured.

The analogy seems to work only if we assume that the relevant positive causal process is the braking process. But this seems to be an implausible choice. First, with regard to the effect in question (P's being harmed) braking seems to be described more appropriately as a negative causal process (taking away the kinetic energy of the car that threatens P). Second, there does not even seem to be a braking process because D did not initiate one. The more relevant positive causal process seems to be the car's approaching P. But this process is not even interrupted. In fact, the case seems to involve overdetermined double prevention: the braking process would have prevented the accident if it had not been prevented by D's failure to brake and the brakes' not being functional. The crucial disanalogy, thus, is the following: in the fielder cases there are two factors that compete in preventing an outcome (the broken window). In the braking case the outcome is not even prevented (P is injured) because there are two factors that compete in preventing the effect that would have prevented the outcome. The most natural choice of an actual cause in the braking case, thus, seems to be the fact that the car approaches P (or was steered in the direction of P).

The legal case is concerned with the braking process. So let us see whether the analogy to preemptive prevention can still be used in order to shed light on this. The appropriateness of the causal model given above depends, among other things,

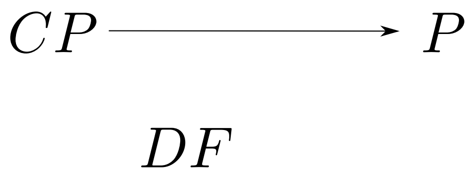

Figure 6.4.: A more appropriate representation of the braking case.

upon our willingness to allow the combination $DF = 0 \wedge BF = 0$. This is the scenario where D does not fail to operate the brakes and the brakes do not fail and, as a result, P is not injured. But this scenario seems to be highly unlikely because the case description specifies that the brakes are not functional. In fact, the braking case seems to be more similar to the fielder-wall scenario than the fielder-fielder scenario. Wright acknowledges the similarity with the fielder-wall scenario and argues that the "defects in the braking system [...] are like the wall in the thrown ball example" (320).

Following the argument that I have developed in the foregoing section, it would then be appropriate to represent the case with a model that does not allow the combination $DF = 0 \wedge BF = 0$. Here is such a model (see figure 6.4). $CP = 1$ shall again represent the fact that the car approaches P. In the given case D fails to operate the brakes while approaching P. But the brakes' not being functional is a structural reason for P being independent of DF. The fact that the brakes are not functional is represented by the fact that there is no directed edge between DF and P. However, there is a directed edge between CP and P, reflecting the fact that P's being harmed depends on whether the car approaches P.

According to this model, DF is not an actual cause of $P = 0$. Neither is the brake's not being functional an actual cause. In fact, the only actual cause is the fact that the car approaches P as argued above. However, Wright argues that D's failure to brake *is* an actual cause of P's being injured and he aims to support this by claiming that the fielder is an actual cause in the fielder-wall scenario (see Wright (2011), 316). That is, Wright bites the bullet and accepts these counter-intuitive claims that causal models avoid if the model is adjusted in an appropriate way.

This seems to be an undesirable result that raises doubts regarding the NESS criterion as a suited criterion for actual causation. But the problem is even worse because Wright needs to commit to the claim that the fielder is an actual cause even in the fielder-gravitation case, or so I shall argue in the following.

Wright argues that his approach does not entail that the fielder is an actual cause in the fielder-gravitation case:

> "Unlike the first two versions, Collins is assuming that the ball lacked sufficient speed when it was released to reach the stated end point. The lack of sufficient speed when it was released caused the failure of the positive causal process of the ball's reaching that very distant point the instant the ball was released and thus pre-empted the potential negative causal effect on that process of [the fielder's] subsequent catching of the ball" (2011, 317).

According to Wright, the actual cause of the ball's not hitting the distant comet supposedly is associated with Earth's gravitational field. The reason is that the ball does not have sufficient kinetic energy to evade Earth's attractive force. And this is the case at the very beginning of the process, that is, before the fielder catches the ball. In other words, with regard to the fielder-gravitation case Wright seems to accept that the gravitational field is a relevant constraint of the ball's possible trajectories.

This reply seems to be exactly right. But from the perspective of Wright's principled account there is a problem: why doesn't this work as a response to the fielder-wall scenario as well? From the very beginning of the process the ball does not have the kinetic energy to get through the wall and destroy the window (the ball is not a cannonball with sufficient momentum to destroy the wall and the window behind it). But this does not make the wall the actual cause of the window's being saved, at least according to Wright.

Here is one way Wright might want to respond: what really matters is *when* the respective forces act on the ball. Take the ball to be travelling in the direction of the distant comet. If no forces were acting on the ball, the ball would approach the distant comet with uniform velocity, as described by Newton's first law of motion. Yet, there are forces acting on the ball and they do so in temporal order: first Earth's gravitational attraction together with atmospheric friction and then the force of the fielder's hand. Thus, gravitation *does* come first and, thus, preempts the fielder.

The problem with this anticipated response is the last "thus." If temporal order of forces on the ball is so important, then we should be precise. It is true that the gravitational force acts on the ball first, but only up to the point where the ball is caught by the fielder. Suppose the Earth's gravitational attraction would end just behind the fielder's hand (and other forces, like friction would also be eliminated). If the fielder had not caught the ball under these circumstances, then the ball would have continued on its trajectory until its impact on the distant comet. Thus, under these circumstances it *is* quite plausible to view the fielder's action as the (but-for) cause of the ball's not hitting the comet. Conversely, the reason why we do not think of the fielder under normal circumstances as the cause is that behind the fielder Earth's gravitational field does not end.

Thus—presuming that the analogy works—Wright's account seems to face a dilemma. One option for Wright is to claim that the failure to operate the brakes is an actual cause and the same holds for the first fielder in the fielder-fielder case and for the fielder in the fielder-wall case. But then Wright is also committed to identifying the fielder in the fielder-gravitation case as an actual cause. The other option is to reject that the fielder is an actual cause in the fielder-gravitation case. But then Wright must accept that the fielder is not an actual cause in the other scenarios either and that the failure to operate the brakes is not an actual cause.

Either position is implausible. Thus, there must be other considerations that legitimately influence our judgements of actual causation in these cases. The analysis given here suggests that these considerations concern the question which scenarios we should be taking seriously. It is a reasonable scenario that upon failure of the first fielder the second fielder also fails to catch the ball. Therefore, in the actual scenario we identify the fielder as a cause of the window's not being hit. This is different in the fielder-wall and the fielder-gravitation case. These do not involve reasonable scenarios where the ball hits the window or the distant comet. Thus, in these cases the fielder is not an actual cause. Neither is the wall or the gravitational field an actual cause in these scenarios. Instead, they feature in (causal) explanations of why the ball-catching process is independent of the window's or comet's not being hit.

Analogously, the braking case does not involve a reasonable scenario in which the car is stopped. Thus, D's failure to brake is not an actual cause of P's being injured. Neither is the defective brake an actual cause, even though it features as a part of the (causal) explanation why the car could not have been stopped. Does this mean that P's being harmed (like the window's remaining intact) has no actual cause at all? No. This is where the analogy between the fielder cases and the braking case breaks down. P's being injured is caused by the car that approaches P.

6.7. Conclusion

In this chapter I have explored consequences of the foregoing discussions for the notion of actual causation in the law. I have argued that my distinction between two kinds of context-sensitivity sheds new light on the feasibility of a principled (context-independent) approach to actual causation. Wright's framework suggests a way to separate context-sensitivity$_1$ (which is related to issues of causal selection) from the causal part of the legal inquiry. Context-sensitivity$_2$ (which is related to taking seriously certain combinations of variables and values that violate a model's structural equations), however, is not so easily separated from the causal inquiry.

7. Causation and the Problem of Disagreement

In Chapter 3 we have seen that the original motivation for incorporating the default/deviant distinction is derived from the Problem of Isomorphism.[1] The idea was that defaults help explain why different causal judgements apply to pairs of target systems even if the systems supposedly have isomorphic causal models.[2] However, we have also seen that the explanation of causal judgements based on the default/deviant distinction is not entirely satisfactory. The explanation involves *ad hoc* assumptions about what is taken to be normal or abnormal. Moreover, in the discussed instances the default/deviant distinction is not even necessary for solving the Problem of Isomorphism. By choosing the causal models more carefully we could show that the pairs of cases do not have isomorphic structure. Moreovoer, the more carefully chosen models help account for the different causal judgements that apply to the different target systems.

In a recent article Thomas Blanchard and Jonathan Schaffer (2017) suggest a generalization of this strategy, which I shall call the *adjust-the-model argument*. They argue that causal reasoners should take the following to be a useful heuristic with

[1] Chapter 7 is an extended version of my article "Causation and the Problem of Disagreement" (Fischer (forthcoming b)) accepted for publication by *Philosophy of Science* on 03/30/2021 (`https://doi.org/10.1086/714852`).

[2] In the following the term 'causal model' will refer to standard causal models, that is, models without the distinction between default and deviant values. Models with defaults will be called extended causal models.

regard to the Problem of Isomorphism: "*When confronted with structurally isomorphic but causally distinct cases, suspect that at least one of the models is impoverished or otherwise non-apt*" (205). They also argue that defaults "come close to a free parameter in an otherwise so precise and objectively constrained formalism, which basically gives the theorist leeway to hand-write the result she wants" (192). Thus, according to Blanchard and Schaffer, the default/deviant distinction does more damage than good to the formalism of causal models.

In this chapter I shall provide a more nuanced account of the benefits of the default/deviant distinction. The account will be based on an analysis of the adjust-the-model argument. In particular, I shall argue *pace* Blanchard and Schaffer, that there are situations where the default/deviant distinction is a useful supplement to causal models. I shall grant that Blanchard and Schaffer's criticism of defaults as a solution to the Problem of Isomorphism is right. However, there is another far less prominent problem: the Problem of Disagreement. And I will show that this problem gives rise to a genuinely new argument for incorporating the default/deviant distinction.

The Problem of Disagreement has first been introduced by Halpern and Hitchcock (2015). It arises from cases where agents disagree in their causal judgement even though they base their judgement on the same assumptions about the underlying causal model. The Problem of Disagreement is related to well-known examples of disagreement over what is 'the cause' of a given effect, discussed, for example by Collingwood (1938), Malone (1956), Hanson (1958), and van Fraassen (1980). The main difference is that the Problem of Disagreement involves the explicit assumption that the disagreeing agents base their causal claims on the same underlying causal model. As in the Problem of Isomorphism, Halpern and Hitchcock take this to indicate that the agents' causal judgements depend not only on assumptions about causal structure but also on a distinction between default and deviant behaviour.

I will show that this argument allows two possible readings. First, it can be read as involving descriptive claims about how agents *do* reason about causal models in contexts where they disagree. This reading seems to be vulnerable to a version of Blanchard and Schaffer's adjust-the-model argument. If two agents disagree about judgements of actual causation with regard to a particular situation, we should expect that these agents also disagree about the underlying causal model. Second, the argument can be read as involving prescriptive claims about how agents *should* reason about causes when they disagree. Here the adjust-the-model argument does not apply. I will argue that it would be wrong to require that the agents support their conflicting causal judgements with different models. Instead, I will argue, causal models should be understood as a representative tool that helps express causal claims that go beyond causal judgements that are based on potentially idiosyncratic normative presumptions. If understood in this way, they can help resolve disagreement over causes by giving a framework for disentangling normative and epistemic dimensions of disagreement. And this function can only be fulfilled if models incorporate the default/deviant distinction. I will illustrate this claim with an example that concerns the causal role of Search and Rescue missions in the Central Mediterranean with regard to increasing numbers of deaths through shipwreck in 2015 and 2016.

In section 7.1 I will have a closer look at Blanchard and Schaffer's case against the Problem of Isomorphism. In particular, I shall introduce in more detail the adjust-the-model argument as one of three challenges that Blanchard and Schaffer raise against proponents of the default/deviant distinction. In section 7.2 I will introduce the Problem of Disagreement. I shall briefly introduce Halpern and Hitchcock's main example which concerns disagreement about the causal status of omissions. In section 7.3 I will argue why Halpern and Hitchcock's example is not a convincing case for defaults. More specifically, I will point out that this version of the Problem of Disagreement is vulnerable to a version of Blanchard and Schaffer's adjust-the-model argument: those who disagree about causal judgements tend to disagree about the causal model as well. In section 7.4 I shall take a step back and examine the function of extended causal models. In sections 7.5 and 7.6 I will argue that extended causal models can help us disentangle disagreement that arises for normative reasons from disagreement that arises for epistemic reasons. I will illustrate this point by drawing from a case of actual disagreement over causes. This example concerns the causal role of Search and Rescue missions in the Central Mediterranen performed by non-governmental organisations in the context of increasing numbers of deaths in 2015 and 2016.

7.1. The Adjust-the-Model Argument

Blanchard and Schaffer put forward three main lines of criticism against incorporating the default/deviant distinction. The first line of criticism, the adjust-the-model argument, will be the focus of the following discussion but it will be useful to have all three objections on the table before we start.

First, according to Blanchard and Schaffer, the default/deviant distinction is unnecessary. The underlying argument is a generalization of the observations discussed in Chapter 3. There we have seen that two prominent instances of the Problem of Isomorphism arise only because one of the involved models did not provide an appropriate representation of the underlying target system. Blanchard and Schaffer argue that incorporating the default/deviant distinction is not a conclusion supported by the Problem of Isomorphism. Instead,

> "[t]he right moral is to dump at least one of the two models invoked on the grounds that it fails to be apt. Indeed it seems to us that the follow-

> ing is a good heuristic: *When confronted with structurally isomorphic but causally distinct cases, suspect that at least one of the models is impoverished or otherwise non-apt.* [...] [T]his heuristic functions as a useful 'warning signal' for the theorist that some non-apt model may be in use, which may trigger her to check both models more closely with her independently developed aptness constraints" (205f, emphasis original).

The relevant aptness constraints are rules for selecting a set of variables $\mathcal{V}$ that constitutes the causal model (see section 3.2). Blanchard and Schaffer focus on three such rules. First, the "variables should represent enough events to capture the essential structure of the situation being modelled", second, "[a]dding additional variables should not overturn the causal verdicts" (183) and, third, "variables should not be allotted values that we are not willing to take seriously" (182).

The first two rules are the rules that they take the simple model of bogus prevention to violate. The simple model is impoverished because it does not reflect the structurally essential fact that no neutralization took place. Adding a corresponding variable overturns the verdict that the actions of the bodyguard are an actual cause (see section 3.8). Blanchard and Schaffer take the third rule to help us with cases like the gardener/queen example: some flowers would not have died if either the gardener or the Queen of England had watered them and it needs to be explained why we tend to identify only the gardener as an actual cause.[3]

> "It is because we are willing to indulge in the fantasy of the gardener watering the flowers [...], but just can't imagine the queen stooping to the job, that we feel an asymmetry. If so then [the constraint to represent only serious possibilities]—which does independent work—was all we needed to explain the gardener/queen asymmetry. There is no apt causal

[3] In the gardener/queen case the problem arises from a symmetry that is internal to the model, not from two causal models that have isomorphic structure.

> model in which wiggling whether the queen waters the flowers wiggles the fate of the flowers, because there is no apt causal model that considers so ridiculous a scenario as the queen of England popping by, watering can in hand, to engage in random acts of gardening" (197).

Figure 7.1A gives a representation of the gardener/queen case that Blanchard and Schaffer consider to be problematic. Blanchard and Schaffer think that this is not an apt model because $Q = 1$ represents a scenario that we are not willing to take seriously. Thus, they suggest eliminating variable Q which leads to the simpler model in figure 7.1B. This model reproduces the plausible verdict that only the gardener is an actual cause of the flowers' death.

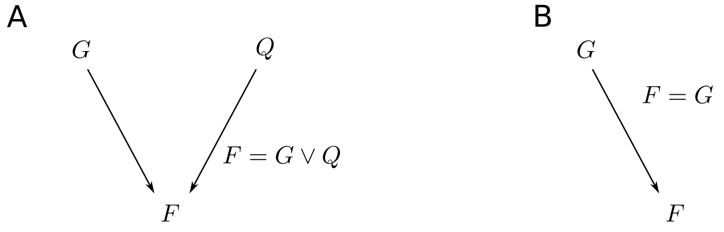

Figure 7.1.: Employing the adjust-the-model strategy for solving the gardener/queen case. A: The flowers survive ($F = 1$) if either the gardener ($G = 1$) or the queen ($Q = 1$) waters the flowers. B: The flowers survive if and only if the gardener waters them.

Blanchard and Schaffer argue that there are two ways to understand this argument. First, the aptness constraint can be understood as reflecting a metaphysical asymmetry between the gardener and the queen. An alternative view is that the underlying asymmetry is merely psychological. According to this understanding, both the gardener and the queen are causes of the flowers' death but there are psychological reasons for our focussing on the gardener. According to this view, the constraint to represent only serious possibilities "may be interpreted not as an aptness constraint on models, but as a descriptive psychological claim about which causal models are most readily available to us when we form our causal judgements" (198).

Blanchard and Schaffer's second line of criticism is that the default/deviant distinction involves unclarities. Most proponents of the default/deviant distinction relate it to an underlying theory of typicality and normality that involves a range of possibly conflicting standards. This can lead to problems as, for example, in "[m]ost people speed. If the posted speed limit is 55 miles per hour, is driving at 55mph normal for conforming to the law, or abnormal for violating the statistical expectation?" (193). The worry underlying this point is that the unclarity associated with the default/deviant distinction stands in contrast with the precise theoretical framework of standard causal models. The authors reject the idea that we should supplement the "precise and objectively constrained formalism" with "a free parameter [...] which basically gives the theorist leeway to hand-write the result she wants" (192).

The third line of criticism is that incorporating the default/deviant distinction is psychologically implausible. Proponents of defaults assume that the results from empirical studies on causal reasoning such as those based on the the pen vignette (Knobe and Fraser (2008), see Chapter 2) reflect judgements that arise from the competent use of a norm-laden notion of actual causation. But, according to Blanchard and Schaffer, such causal judgements are rather to be explained by biases that are associated with certain heuristics that support our use of a norm-free causal notion. In the case of the pen vignette, for example, test subjects supposedly should have identified both Professor Smith *and* the administrative assistant as actual causes. The fact that test subjects ascribe a higher relevance to Professor Smith is to be explained by norm-related presuppositions that interfere with the correct use of a norm-free notion of actual cause. Blanchard and Schaffer support this claim with an analogy to probability judgements. Test subjects tend to overestimate the probability of a car accident after being presented with dramatic images of such accidents (Tversky and Kahneman, 1973; Kahneman and Miller, 1986). Blanchard and Schaffer argue that, in analogy to the pro-default argument, we would have to presume that the test subjects make competent use of a norm-laden notion of probability. This is of course implausible. Instead the result is to be explained by a norm-free notion of probability the use of which is guided by a heuristic. The heuristic is affected by a bias that results from the test subjects' being presented with the dramatic pictures.

In the following I will focus on Blanchard and Schaffer's first line of criticism, which I shall refer to as the *adjust-the-model* argument. While I grant that this is a strong argument against the Problem of Isomorphism, I will show that in the context of the Problem of Disagreement we need a more nuanced account.

There is an important tension between Blanchard and Schaffer's three arguments. Suppose I am a proponent of the idiosyncratic (and potentially biased) view that the queen is in charge of watering the flowers in the municipal gardens and that the gardener for some reason is not supposed to water them.[4] According to the adjust-the-model strategy, I am supposed to represent only those scenarios that I take to be serious possibilities. Thus, I will end up with a model in which variable Q is the only cause of variable F. But this is a problem. Because now my idiosyncratic view does not only spoil my judgements of actual causation, but also the corresponding causal model!

The underlying point is the following. Blanchard and Schaffer argue that the default/deviant distinction is unclear and reflects biases. Then they suggest to solve cases like the gardener/queen example by adjusting the models on the basis of considerations about what scenarios are to be taken seriously. But what is a scenario that is to be taken seriously? Presumably this depends on ideas related to normality—otherwise it would be easy to generate counterexamples to the strategy.

[4] Blanchard and Schaffer construct a similar case, where the gardener is a member of a secret society that does not allow her to water inedible plants.

But this means that the constraint on models is no less unclear than the criteria for the default/deviant distinction. It seems like we haven't gained anything by shifting the problem of unclarity from the default/deviant distinction to the criteria for selecting a suitable set of variables $\mathcal{V}$. In fact, exploiting the serious possibility rule as a constraint on $\mathcal{V}$ makes the problem even worse. For now the unclarities are not confined to the defaults but they infect the whole model.

My argument in this chapter is that there are situations where normality considerations should not affect the choice of variables in $\mathcal{V}$. If there is unclarity associated with norms, then defaults are a better place for it.

7.2. The Problem of Disagreement

Consider the following case of causation by omission, which is a variant of the gardener/queen case and which is provided by Halpern and Hitchcock:

> "[W]hile a homeowner is on a vacation, the weather is hot and dry, her next-door neighbour does not water her flowers, and the flowers die. Had the weather been different, or had her next-door neighbour watered the flowers, they would not have died" (414f).

Halpern and Hitchcock argue that since the flowers' death depends counterfactually on both the weather and the neighbour's omission it seems like a counterfactual theory of causation cannot distinguish between these factors. However, according to some authors (e.g. Beebee (2004); Moore (2009)), the weather is a cause of the flowers' death but not the neighbour's omission to water them. Halpern and Hitchcock flag this as the "problem of isomorphism." Note that (as in the gardener/queen case) this is a somewhat non-standard use of the term "problem of isomorphism." In Chapter 3 we have seen that the Problem of Isomorphism typically is taken to arise from pairs of cases that are represented by isomorphic causal models. Here there is only one case and one model at stake and the problem arises from a symmetry that is internal to the model: there are two factors that stand in the same kind of structural relation to the effect, but only one is identified as actual cause.

According to Halpern and Hitchcock, there is "an even deeper problem. There is actually a range of different opinions in the literature about whether to count the neighbour's negligence as an actual cause of the flowers' death [...]. *Prima facie*, it does not seem that any theory of actual causation can respect all of these judgments without lapsing into inconsistency" (415). This, according to Halpern and Hitchcock, is the Problem of Disagreement.

The Problem of Disagreement arises where the following two conditions hold. First, there are two (or more) agents A_i that have conflicting judgements of actual causation with regard to the same target system. For example, theorists like Beebee and Moore argue that only the weather is an actual cause because they think that omissions cannot be actual causes. They disagree with theorists like Lewis (2000; 2004) and Schaffer (2000b; 2004) who think that the neighbour's negligence is also an actual cause because they think that omissions are genuine causes. Second, it has to be the case that these opposing agents agree on the underlying causal model M. In the flower case Halpern and Hitchcock take this to be a model consisting of the following three variables (Halpern and Hitchcock, 2015, 437). First, $H = 1$ if the weather is hot and dry and $H = 0$ otherwise. Second, $W = 1$ if the neighbour waters the flowers, and $W = 0$ otherwise. Third, $D = 1$ if the flowers die, and $D = 0$ otherwise. The flowers die if the weather is hot and the neighbour fails to water them: $D = H \land \neg W$.

The Problem of Disagreement is related to well known examples of disagreement over what is 'the cause' of a given effect, discussed, for example by Collingwood (1938), Malone (1956), Hanson (1958), and van Fraassen (1980). For example, Hanson describes a case where "the cause of death might have been set out by a physician as 'multiple haemorrage', by the barrister as 'negligence on the part of the driver', by a carriage-builder as 'a defect in the brakeblock construction', by a civic planner as 'the presence of tall shrubbery at that turning'" (Hanson, 1958, 54). These authors emphasize that what counts as 'the cause' depends on criteria that are contextual. One way context-dependence plays out is when different agents make causal claims about the same situation but with different background assumptions or pragmatic aims in mind. The main difference between these earlier discussions and Halpern and Hitchcock's more recent treatment is that the Problem of Disagreement involves the explicit assumption that the disagreeing agents base their causal claims on the same underlying causal model.

When there is an instance of the Problem of Disagreement, what exactly do the agents disagree about? According to Halpern and Hitchcock, the disagreement is about the actual cause of the outcome. That is, what is an actual cause according to one agent is not an actual cause according to another agent. But wouldn't this imply an implausible metaphysical view according to which causation is subjective? Halpern and Hitchcock argue to the contrary. Actual causation is taken to be a subjective and context-dependent notion that is to be distinguished from an underlying and objective notion of causal structure.

Going back to Blanchard and Schaffer's third line of criticism one might think that the disagreement appears to affect only idiosyncratic biases and, thus, should not be part of a theory of the notion of causation. One might think that the term 'causation' should be reserved for the underlying structure and explain the rest by biases that are associated with certain heuristics. But there seems to be a disanalogy to Blanchard and Schaffer's case about the notion of probability and probabilistic reasoning. For an illustration take the conjunction fallacy, also known as the Linda problem (Tversky and Kahneman, 1983). The Linda problem arises from a situation where test subjects are presented with a description of Linda as an outspoken, bright, and politically engaged person who has majored in philosophy. These test subjects tend to say that it is more likely that Linda is a bank teller and active in the feminist movement than that she is bank teller. This is clearly false since the set of feminist bank tellers is a subset of all bank tellers. The crucial point here is that if you are not familiar with the problem, it is easy to give the wrong estimation. But once you know that with the estimation you committed the conjunction fallacy, you see immediately that you have been led astray. In particular, there is no reason to stick to your initial judgement.

But this is different in the case of judgements of actual causation. In Chapters 4 and 5 I have argued that judgements of actual causation and in particular the norm-dependent judgements that are relevant in the Problem of Disagreement do not express a mere bias but fulfil certain functions. In particular they identify suitable targets of intervention or indicate who is responsible for some outcome. Disagreement about these issues will not be resolved by merely pointing out that on the level of causal structure everyone agrees. Does this mean that the disagreement concerns suitable targets of intervention and responsibility rather than causation? If 'causation' means causal structure, then this seems exactly right. But this is not what the norm-dependent notion of actual causation is taken to refer to.

Next, let us see how Halpern and Hitchcock aim to resolve this instance of the Problem of Disagreement by employing the default/deviant distinction. As we have seen in Chapter 3, Halpern and Hitchcock take the default/deviant distinction to be one that concerns the values of particular variables. If all variables in a causal model take on particular values, then the model represents a possible world. Moreover, worlds can (often but not always) be compared on a normality scale, where the normality of a world is a function of the number of variables that take on their default values.

Halpern and Hitchcock argue that "[t]hose who maintain that omissions are never causes can be understood as having a normality ranking where absences or omissions are more typical than positive events" and Halpern and Hitchcock take this to reflect "a certain metaphysical view: there is a fundamental distinction between positive events and mere absences, and in the context of causal attribution, absences are always considered typical for candidate causes" (437f). This, according to Halpern and Hitchcock gives rise to the following normality ordering:

$$(H = 0, W = 0, D = 0) \succ (H = 1, W = 0, D = 1) \succ (H = 1, W = 1, D = 0)$$

Here the most normal world is the world where the weather is not hot and dry and the neighbour does not have to water the flowers in order to save them. The actual world is less normal because the weather variable takes on a non-default variable (and the flowers die). This is taken to be more normal than the world where the flowers are being watered and survive as a result. With this normality ordering the norm-sensitive definition of actual causation identifies the weather as an actual cause but not the neighbour's negligence. This is because the normality criterion of actual causation discussed in Chapter 3 allows as a witness only the world where $(H = 0, W = 0, D = 0)$.

By contrast, an advocate of the view that omissions are always causes can be understood as subscribing to the following normality ordering:

$$(H = 0, W = 0, D = 0) \equiv (H = 1, W = 1, D = 0) \succeq (H = 1, W = 0, D = 1)$$

Here the two worlds where the flowers do not die are equally normal and they are taken to be at least as normal as the world where the flowers die. Consequently, both the weather and the neighbour's negligence fulfil the normality criterion and qualify as actual causes of the flowers' death.

In conclusion, the Problem of Disagreement is supposed to show that causal judgements track considerations that go beyond the content of causal models. As in the Problem of Isomorphism the idea is that the problem can be solved by incorporating a default/deviant distinction.

7.3. Problems with the Problem of Disagreement

There are two problems with Halpern and Hitchcock's account of the Problem of Disagreement. First, it seems quite unlikely that Beebee and Moore would agree with the claim that absences or omissions are generally more normal than positive events. In fact, according to each of the many dimensions of normality, there seem to be clear counterexamples. Living humans more frequently breathe than not, functional alarm clocks go off, we are legally and morally required to help those whose lives are threatened through an accident. The kind of metaphysical point that Beebee and Moore make with regard to the causal status of omissions seems to be independent of claims regarding the normality of omissions. Thus, it seems Halpern and Hitchcock have chosen an example where defaults do not do the explanatory work that they expect them to do.

Second, suppose for the sake of the argument that there is an agent who subscribes to the view that absences are always considered typical. These assumptions about typicality will most likely also influence what an agent takes to be a far-fetched or a serious possibility—considerations that we should expect to affect the agent's preferred causal model. But if this is the case, then this agent disagrees with the proponent of absences as causes already at the level of the standard causal model.

So, if we take the Problem of Disagreement to give rise to an argument for defaults, then it seems like this argument faces the same kinds of difficulties as the argument from the Problem of Isomorphism. In particular, it seems like there is not really a problem in the first place if we choose what seem to be the most plausible models.

The fact that we have to expect agents to agree already at the level of causal models poses a general problem for the Problem of Disagreement. The claim that there are agents who disagree about actual causes but agree on the underlying causal model seems to involve implausible empirical assumptions about the involved agent's sets of believes.[5]

7.4. What is the Function of Extended Causal Models?

In the following sections I will argue that there is an alternative reading of the Problem of Disagreement. This reading of the argument does not rely on descriptive claims about how disagreeing agents *do* use extended causal models to support their reasoning. Instead it focuses on prescriptive claims about how disagreeing agents *should* use extended causal models. I will argue that the alternative reading shows that in some cases the default/deviant distinction is a useful extension to standard causal models.

Let us begin by considering what the function of extended causal models is, in Halpern and Hitchcock's framework. When Halpern and Hitchcock describe the default/deviant distinction as a conservative extension, they "envision a kind of conceptual division of labour where the causal model $(\mathcal{S}, \mathcal{F})$ represents the objective patterns of dependence that could in principle be tested by intervening on the system, and $\geq$ represents the various normative and contextual factors that also influence judgments of actual causation" (435). So, at a first glance it looks like causal reasoning involves considerations that are located at two distinct levels. First, there is the level of standard causal models. These represent the objective patterns of counterfactual dependence. Second, there is the level of judgements

[5] Phillips and Cushman (2017) provide evidence to the effect that moral norms constrain an agent's representation of what is possible. The authors show that under time pressure test subjects are less likely to judge it possible to act immorally or irrationally. If agents do not consider such actions possible, it seems plausible to assume that they would not include them into a causal model of such a situation.

of actual causation. These judgements are influenced by the normality ordering $\geq$ which reflects normative and contextual considerations.

However, at a second glance the conceptual division of labour does not seem to work as straightforwardly. First, Halpern and Hitchcock point out that objectivity on the level of standard causal models means that "once a suitable set of variables has been chosen, there is an objectively correct set of structural equations among those variables" (431f). Thus, the causal model $(\mathcal{S}, \mathcal{F})$ itself is not objective. For the choice of the signature $\mathcal{S}$ (the choice of relevant variables and their possible values) is likely to be governed by criteria that are sensitive to normative and contextual factors.

Second, even the judgements of actual causation need to have some objective core. Otherwise they could hardly help us "identify appropriate targets of corrective intervention" (432). Take, for example, the claim that Professor Smith is the actual cause of the receptionist's problem of pen shortage. This claim is clearly context-sensitive in the sense that it depends, for example, on the department's pen policy. But it also relies on a relation of counterfactual dependence that holds independently of this norm.

But if causal models (plus information about the variables' actual values) and judgements of actual causation are so similar, why do we need both? Couldn't we just make do with either one of them? The function of claims of actual causation has been the topic of Chapter 5. Knowing, for example, that Professor Smith is the actual cause of the pen shortage, indicates that Professor Smith is to be blamed and that her behaviour is a suitable target for avoiding future problems of pen shortage. Claims of actual causation are highly selective. And this has the advantage that they can guide agency very straightforwardly—we do not need to consider further dependencies such as the one concerning the administrative assistant. Presumably, causal models are somewhat closer to the objective structure because they allow representing larger chunks of it. They express complex counterfactual dependencies that are not captured by a simple claim of the form '$X = x$ is an actual cause of $Y = y$.' These larger chunks still depend upon norms, but do so to a lesser degree because selection does not have to be constrained so narrowly.

In the following I shall argue that the Problem of Disagreement helps to indicate one distinctive advantage of causal models: causal models can help us provide a representation of disagreement about causes that is more conducive to resolving the disagreement than the bare claims of actual causation. Moreover, I shall argue that this function is sometimes (but not always) crucially facilitated by incorporating the default/deviant distinction.

7.5. An Example: Search and Rescue Missions

Let me introduce an example that shall serve as an illustration. The example concerns the causal role of Non-governmental Search and Rescue missions (NGO SARs) for fatalities of refugees in the Central Mediterranean. According to Frontex,[6] the European Border Control Agency, NGO SARs are an actual cause of the increase of the number of deaths in the Central Mediterranean in the period from 2015 to 2016. Their narrative of NGO SARs as acting like a "pull factor" has been taken up by a number of organizations and politicians across Europe. On the other hand, it has been argued that NGO SARs are only one factor acting within a complex causal structure, and that it is erroneous to describe NGO SARs as *the* cause of the increase. In particular, I will look at a study performed by Forensic Oceanography[7] and show that the most natural way to understand their criticism of Frontex's claim is to see it as an attack on Frontex's assumptions about the causal model.

Let us begin with a closer look at the claims put forward in the Frontex report.

[6] The following is based on claims made in the risk analysis report for 2017 (FRONTEX (2017)).

[7] Forensic Oceanography is part of the Forensic Architecture agency which is specialized on investigating violations of human rights and which is located at Goldsmiths, University of London.

The report describes an increase of the number of deaths of refugees in the Central Mediterranean and states the following.

> "In this context it transpired that both border surveillance and SAR missions close to, or within, the 12-mile territorial waters of Libya have unintended consequences. Namely, they influence smugglers' planning and act as a pull factor that compounds the difficulties inherent in border control and saving lives at sea. Dangerous crossings on unseaworthy and overloaded vessels were organised with the main purpose of being detected by EUNAVFOR Med/Frontex and NGO vessels. Apparently, all parties involved in SAR operations in the Central Mediterranean unintentionally help criminals achieve their objectives at minimum cost, strengthen their business model by increasing the chances of success. Migrants and refugees – encouraged by the stories of those who had successfully made it in the past – attempt the dangerous crossing since they are aware of and rely on humanitarian assistance to reach the EU" (FRONTEX, 2017, 32).

According to this passage, the presence of SARs (both NGO and state-led operations) near the Libyan coastline has two effects. First, they influence the smugglers' strategies in a way that makes the crossing more risky. The claim is that the presence of SARs gives a sense of security that allows smugglers to offer crossings on vessels that are unseaworthy and overloaded. Second, they lead to an overall increase in attempted crossings. Again the claim is that the presence of SARs gives a sense of security that encourages more migrants and refugees to risk their lives.

The report goes on, stating that "[c]losely related issues are the safety of migrants and refugees and, most significantly, the increasing number of fatalities" (32). After reporting estimates of the fatalities in 2016 the report states that "[t]he increasing number of migrant deaths, despite the enhanced EUNAVFOR Med/Frontex surveillance and NGO rescue efforts, seems paradoxical at first glance" (33). But then the report relates the increase of fatalities to a change in the smugglers' tactics: "[t]he rising death toll mainly results from criminal activities aimed at making profit through the provision of smuggling services at any cost" (33).

It seems fair to assume that the above quoted passages can be summarized by the following causal model, consisting of three variables (see figure 7.2A). First, S shall represent the presence of search and rescue missions. Second, C shall be a factor that represents the risk level of the individual crossing and the number of attempted crossings. Third, D shall represent the number of deaths. It is claimed that an increase in S leads to an increase in C, that an increase in C leads to an increase in D, and that an increase in S also leads to a direct decrease in D. The narrative does not allow a more detailed quantification of these functional relations. But there is a possible reading according to which the narrative states that the increase of deaths via the route $\langle S, C, D \rangle$ is larger than the decrease via the route $\langle S, D \rangle$.[8]

The Forensic Oceanography report (Heller and Pezzani, 2017) identifies the pull-factor claim as part of a toxic narrative within a "de-legitimisation and criminalisation campaign" directed at non-governmental search and rescue missions. The aim of the report is an empirical assessment of the claims put forward by Frontex. In particular, the report can be understood as challenging the structural relation between S and C as it is stated by the Frontex report. The report appears to suggest adding another variable X feeding into C that explains the increase in risk level and number of attempted crossings from 2015 to 2016 (see figure 7.2B). Here is a brief

[8] A more precise model can be given by choosing a more fine-grained version of variable C such that C_1 refers to the risk level of the individual crossing and C_2 refers to the overall number of attempted crossings. Then S would have a mixed influence on C_1, according to the Frontex report. First, an increase in S would lead to an increase in C_1, which amounts to Frontex's claim that the presence of SARs decreases the risk aversion. Second, an increase in S would lead to a decrease in C_1 because the SARs are there to help. Moreover, there would be causal relations among C_1 and C_2. Increased demand for crossings will lead to less seaworthy vessels being used but more dangerous crossings will also suppress demand. I choose the more coarse-grained variable C in order to keep things as simple as possible.

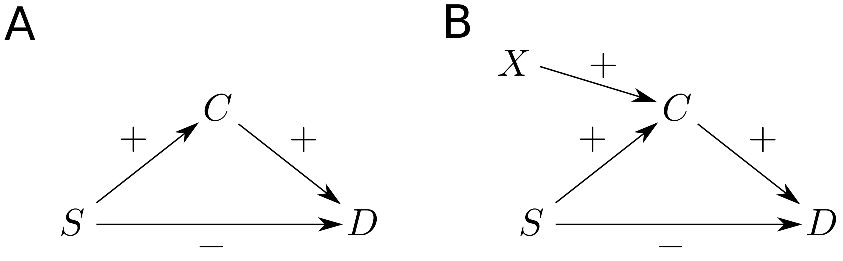

Figure 7.2.: A simple model of the "pull-factor" narrative. A: version of the Frontex report. The increase in risk level associated to indivicaul crossings and number of ttempted crossings *C* is explained through an increase in the presence of Search and Rescue missions *S*. B: version of the Forensic Oceanography report. The increase in *C* is mainly attributed to a change in background factors.

summary of factors included in *X*, according to the Forensic Oceanography report. First, EUNAVFOR Med's[9] mission to counteract smuggling activities in the Central Mediterranean included among other strategies the destruction of smuggler's vessels. As a consequence, smugglers started making use of cheaper and much less seaworthy rubber dinghies. Second, following a shipwreck in 2015 corpses where washed ashore in Zuwara, which was then the main point of departure for migrants at the Libyan coast. As a consequence, the smugglers were pushed out of Zuwara and moved to Sabratha. Sabratha was under the control of militia who then entered the smuggling business helping it to grow to an industrial scale. Third, the options for crossing vary in price. The more expensive options are less risky, for example, because part of the price is paid only after successful landing in Europe. While migrants from the Horn of Africa tended to seal safer deals, the increasing number of poorer migrants from Central and Western Africa could not afford them and chose more risky options. Fourth, the Libyan Coastguard intercepted migrants and thereby, according to the report, complicated the situation at the Libyan coast. In particular, it is not clear to what degree the Libyan Coast Guard collaborated with smugglers or is corrupted. The authors also report instances where the Libyan Coast Guard violently interrupted rescue operations leading to many deaths.

[9] European Union Naval Forces – Mediterranean, also called Operation Sophia.

Moreover, the Forensic Oceanography report describes NGO SARs as a continuation or replacement of preceding state-led search and rescue missions[10]. In particular, the report claims that "[a]iming to deter migrants from crossing the Mediterranean, the EU and its member states pulled back from rescue at sea at the end of 2014, leading to record numbers of deaths. Non-governmental organisations (NGOs) were forced to deploy their own rescue missions in a desperate attempt to fill this gap and reduce casualties." That is, whereas the Frontex report suggests that there is a new kind of search and rescue missions that explains the increase, the Forensic Oceonography report describes the presence of search and rescue activity in the Mediterranean as a default condition.

The report clearly states that "contrary to the claim made by Frontex and others, SAR NGOs have made the crossing safer." Yet, at the same time the Forensic Oceanography emphasizes that

> "[t]he risk that their presence would keep reinforcing the trends we have discussed, thus resulting in effects that are the exact opposite of their humanitarian aims, is real and should not be underestimated. SAR NGOs are acutely aware of this difficult position. As the authors of an internal MSF position document "Unsafe passage" noted: "We are caught in a vicious circle because both smugglers and border guards are exploiting our presence at sea and people continue to die, despite our actions"" (Heller and Pezzani, 2017).

That is, even though the Forensic Oceanography report disagrees with the causal relevance that Frontex ascribes to the negative effects of SARs, it does not deny that there is a causal relation.

[10] See also their earlier report "Death by Rescue" (2016).

7.6. The Role of Defaults

With the example of Search and Rescue Missions at hand let us return to the argument for extended causal models. The point I shall make in this section is that there are cases where extended causal models facilitate the function of representing disagreement over causes better than causal models that do not incorporate the default/deviant distinction.

The disagreement between the Frontex report and the Forensic Oceanography report concerns whether the presence of SARs led to an increase in the number of deaths in the Central Mediterranean. The underlying causal question that is at stake in this example is: why are refugees willing to risk their lives in an attempt to cross the Mediterranean? The presence of SARs (and stories about how they guarantee safety on sea) is considered to be one factor. However, the presence of SARs is surely not a sufficient criterion for someone to risk one's life. It seems plausible to assume that relevant factors in the decision are among others: (i) the situation in the home country, (ii) the hope for a better life in the EU, (iii) the absence of alternative pathways into the EU (legal pathways, or simply pathways that are not as risky).

Suppose each of these factors corresponds to a variable in a causal model such that a variable describing the willingness of refugees to risk their lives depends upon these variables. The disagreement about the causal role of SARs involves agents that have opposing views about which of these variables represent scenarios that are to be taken seriously—for functional, legal, and moral reasons. And these considerations are associated with considerations about who is to be blamed for the refugees' deaths and specific policy interventions that these agents consider to be feasible or not. For example, there is disagreement about the moral and legal feasibility of cutting back life-saving missions on sea. Correspondingly, these agents will have very different views regarding the actual causes (in a norm-dependent sense) of the willingness of refugees to risk their lives.

How *should* this disagreement be represented in terms of standard or extended causal models? One way would be to require that the involved agents agree on a set of variables $\mathcal{V}$ by including all variables that are at stake in the debate and represent their disagreement on the level of the default/deviant distinction. From a humanitarian perspective, for example, life-saving missions would be the moral and legal default. By contrast, certain opposing agents might want to describe the absence of SARs as the default state. But both kinds of agents would be required to include a variable representing SARs.

Alternatively, one could require the views to be expressed by different standard causal models that reflect the individual views about what scenarios are to be taken seriously. This is what is suggested by the adjust-the-model strategy. The advantage is that such models do not incorporate the default/deviant distinction which is considered unclear. The disadvantage, however, is that now the unclarity occurs in a disagreement about which scenarios are to be represented by the model in the first place.

The problem with this strategy is that it leaves unclear whether agents disagree for normative or for epistemic reasons. Suppose agent A_1 does not include a particular variable X in her standard causal model even though agent A_2 thinks that X is a cause of Y. Does agent A_1 mean to say that a change in X would merely amount to a scenario that is not to be taken seriously? Or does agent A_1 mean to imply that even if X were part of the model, it would not stand in a causal relation with Y? Extended causal models clearly fare better in this kind of context. They provide the formal resources that help the involved agents to point out where disagreement arises for normative reasons and where it arises for epistemic reasons. Agent A_1 would be required to include X into the model and clarify whether she takes Y to be independent of X or merely considers X to represent scenarios that from her particular point of view are highly abnormal.

This is particularly important in cases where it is likely that disagreement arises not only about norms but also about the underlying counterfactual dependencies. The core of Frontex's pull factor claim is the counterfactual dependency of C on S. But this claim is of course difficult to assess directly. It involves non-trivial assumptions about the refugees' dispositions to risk their lives. It is also difficult to assess it in an interventionist fashion. For performing testing interventions on the target system is unfeasible in practice. Instead Frontex supports the pull-factor claim by a comparison of the risk levels in 2015 and 2016 and relates this to an increase of the NGO SAR activity over this period. But this argument is of course valid only if all other potential causes for an increased risk level remain constant over this period. In Frontex's selective causal model it looks like this is the case. A more encompassing model that includes information about the specific situation in Libya and the availability of alternatives, however, suggests that Frontex's claims are unwarranted. In order to warrant the pull-factor claim in the context of such a more encompassing model the Frontex report would have to show that the influence of these other factors is irrelevant.

7.7. Conclusion

In this chapter I have argued that there are contexts in which the default/deviant distinction is a useful extension of the standard formalism of causal models. In particular, I have examined the scope of a particular argument against defaults, the adjust-the-model argument. I have taken the adjust-the-model argument to make a convincing case against the Problem of Isomorphism. Then I have argued that it extends to a descriptive reading of the Problem of Disagreement. Agents who disagree about actual causation are likely to support their claims with diverging causal narratives. If we take such narratives to be indicators for underlying causal model assumptions, then it seems descriptively inadequate to assume that agents who disagree about actual causation still agree on the underlying causal model. However, there is a prescriptive reading of the Problem of Disagreement that provides a strong case for incorporating the default/deviant distinction into causal models. And here the adjust-the-model argument does not apply. In cases of disagreement causal models should act as a representative tool for assumptions about the underlying causal structure that are shared by the involved agents. This helps keeping normative disagreement apart from disagreement about the underlying counterfactual structure. This is particularly important in cases with complex and contested structure such as in the context of claims about the causal role of Search and Rescue missions in the Central Mediterranean.

8. Conclusion and Outlook

My main aims have been, first, to argue that we need to be pluralist with regard to actual causation and, second, to provide an analysis of the context-sensitivity of concepts of actual causation. Part has provided the background for my account. In Chapter 2 I have introduced two central problems for theories of actual causation: the problem of redundancy and the problem of selection. In Chapter 3 I have reviewed how causal models have been employed in a range of attempts to provide a unified approach to these two problems. Part II has developed my pluralist account. In Chapter 4 I have argued from an interventionist perspective that we need to distinguish three concepts of actual causation: total, path-changing, and contributing actual causation. In Chapter 5, I have argued that we also need a pluralist account with regard to the function of these concepts: even though the interventionist approach is largely successful there are some instances of reasoning in terms of actual causation that are better explained from the perspective of responsibility. Part III has explored consequences of the pluralist account with particular regard to the context-sensitivity of actual causation. In Chapter 6 I have employed a distinction between two kinds of context-sensitivity in order to show that attempts to provide a principled approach to actual causation in the law face difficulties. In Chapter 7 I have provided a new argument for incorporating a distinction between context-sensitive default and deviant values into causal models. In this last chapter I shall briefly revisit the main results of each chapter and provide an outlook.

In Chapter 2 I have introduced two challenges for theories of actual causation: the problem of redundancy and the problem of selection. In particular, I have introduced four kinds of redundancy: early preemption, late preemption, symmetrical overdetermination, and trumping. Moreover, I have highlighted the challenges that these forms of redundancy pose for counterfactual accounts. Next, I have turned to the problem of selection, which is aptly captured by what I have called *Mill's challenge*: causal selection is capricious and not justified. I have then discussed two kinds of approaches and I have assessed whether they meet Mill's challenge. Contextual-variable accounts (such as contrastive accounts) provide an explanation of the perceived capriciousness of causal selection: claims in the binary form ("A causes B") are incomplete and as soon as we fix the contextual variables the problem disappears. However, contextual-variable accounts do not provide criteria for reconstructing the content of contextual variables and, thus, do not explain why stating causal claims in binary form is mostly successful and, thus, justified. I have then argued that a promising way to approach this problem is to make use of a concept of normality. Normality is clearly a contextual concept. In different contexts different dimensions of normality apply. There are, for example, statistical, functional, and moral norms. Moreover, what is normal according to each particular kind of norm varies as well. Nevertheless, considerations of normality are often largely stable across contexts and they can be justified.

I have also highlighted an interesting parallel between the two problems. Both problems can be explained away if we accept extremely revisionary notions of causation. If we apply an extremely fine-grained individuation scheme for events, then redundancy can be explained away because, for example, the preempting and the preempted factor no longer compete for causing the same effect event. However, this does not seem to be acceptable because the fine-grained individuation scheme leads to a concept of causation that is far too permissive. If we restrict the notion of cause to the totality of jointly sufficient conditions of an effect, then the problem of selection can be explained away. Yet, again, this does not seem to be acceptable because we rarely use the term cause in such a restrictive way. In either case explaining away the problem seems to come at the price of neglecting a range of context-sensitive but important aspects of reasoning in terms of actual causation that are in need of philosophical elucidation.

In Chapter 3 I have reviewed approaches to a unified concept of actual causation that make use of the formal framework of causal models. First, I have discussed a series of accounts that define actual causation as dependence given that certain variables are being held fixed at their *actual* values. These accounts provide a successful treatment of cases involving early preemption but face problems in cases involving symmetrical overdetermination. The discussion of these accounts prepares my concept of path-changing actual causation as defined in Chapter 4. Then I have addressed accounts that define actual causation as dependence given that certain variables are set to *non-actual* values. These accounts succeed also with respect to symmetrical overdetermination and correspond to my concept of contributing actual causation.

However, both kinds of accounts face a challenge that arises from the Problem of Isomorphism: there are pairs of cases that appear to have isomorphic causal models but different causal judgements apply. This suggests that our causal judgements go beyond the content of standard causal models. I have then turned to approaches to the Problem of Isomorphism that incorporate a distinction between default and deviant values. In particular, I have looked at Halpern and Hitchcock's approach that implements the distinction in combination with normality orderings over possible worlds. This approach promises to account not only for issues related to redundancy but also for the problem of selection. Thus, it seems to provide a unified account of actual causation.

In Chapter 4 I have provided the first part of my pluralist account of actual causation. According to my account, we need to distinguish total, path-changing, and contributing actual causes. The definitions of these concepts are closely related to existing definitions discussed in Chapter 3. However, so far it has not been acknowledged that these definitions describe different and equally legitimate concepts of actual causation. I have provided two lines of argument for this claim. First, I have provided three toy examples that raise problems for unified accounts. The examples have in common that there are two factors that both seem to qualify as actual causes. Yet, at the same time, there is an important asymmetry between these factors. Extant theories face the problem that they either describe both factors as actual causes but then they cannot explain the asymmetry. Alternatively, extant theories dismiss the intuition that both factors are actual causes and account for the asymmetry by identifying only one of the factors as actual cause. My explanation is that each example involves two actual causes but that they are different kinds of actual causes, which explains the perceived asymmetry.

The second line of argument for my pluralist account is a functional argument. Given that the notion of actual causation has the purpose to help us identify suitable targets of intervention, I argue, we need to distinguish different concepts of actual causation. The reason is that total, path-changing, and contributing actual causes enable different kinds of control over their effects. Total actual causation means that the effect can be prevented by an intervention on the actual cause. Contributing actual causation means that other and independent factors have to be targeted as well. Finally, path-changing actual causation describes the special case where a primary intervention on the actual cause needs to be combined with a secondary intervention that counteracts the adverse consequences of the primary intervention.

Moreover, I have argued that we can distinguish two kinds of context-sensitivity. Context-sensitivity$_1$ concerns the normality or typicality of the individual variables' possible values. This kind of context-sensitivity concerns total, path-changing, and contributing actual causation if the concepts are combined with a normality ordering over possible worlds. Context-sensitivity$_2$ concerns considerations regarding possible violations of the model's structural equations. This kind of context-sensitivity additionally affects path-changing actual causation.

In Chapter 5 I have scrutinized a key assumption of the foregoing chapter: that the function of concepts of actual causation is to indicate suitable targets of intervention. I have argued that there are two kinds of cases where interventionist accounts have difficulties to explain the function of relevant causal intuitions. First, interventionists have difficulties to explain our interest in certain selective claims of total actual causation. Hitchcock and Knobe's drug vignette describes a case where we identify a norm-violating factor as actual cause even though this factor is not a suitable target of intervention. Second, in cases of late preemption it is difficult to explain the distinction between path-changing actual causes and preempted factors. Both path-changing actual causes and preempted factors correspond to independent causal processes leading up to the effect and in order to prevent the effect, both causal processes have to be intervened upon in the same sense.

In both instances a better explanation of the corresponding causal intuitions is achieved with reference to the retrospective evaluation of responsibility. The reason why retrospective evaluations of responsibility seem to provide a better explanation is related to an asymmetry between the retrospective evaluation of past actions and the prospective evaluation of future actions. Factors that lie beyond the control of an agent and that are not foreseen by the agent affect the retrospective evaluation but not the prospective evaluation.

In Chapter 6 I have examined the context-sensitivity of actual causation in the law. I have argued that Wright's attempt to employ the NESS criterion in order to provide a principled account of actual causation in legal inquiry faces counterexamples. In order to show this I have employed the distinction between context-sensitivity$_1$ and context-sensitivity$_2$ drawn in Chapter 4. I have argued that even if Wright's three-stage framework of legal inquiry may succeed in avoiding issues related to context-sensitivity$_1$, there still remain issues related to context-sensitivity$_2$. I have illustrated these issues with variations of cases involving preemptive prevention and their relation to a legally relevant case where a car accident occurs because a car approaches a person and the driver fails to use the brakes but the brakes were also defective. I have argued that Wright's account runs into problems because it implies highly implausible causal claims.

By employing causal models I have provided a better account of preemptive prevention cases and of Wright's braking case. In a situation where a fielder catches a ball before it hits the window we identify that fielder as actual cause of the window's remaining intact. This is the case even if there was a second fielder who would have caught the ball if the first fielder had failed. The reason is that in this kind of situation it is not a far-fetched scenario that the ball hits the window (both fielders could fail). In the situation where the second fielder is replaced by a solid brick wall, however, the broken window is not a scenario that is to be taken seriously and, thus, we need to assume a causal model that is structurally different from the one that represents the situation with two fielders. The braking case is similar to the situation that involves the wall. It is a far-fetched scenario that the car suddenly stops if the brakes are not functional. Therefore, we should not identify the driver's failure to operate the brakes as actual cause of the accident. Instead the cause of the car's hitting the person is the fact that the car was steered in the direction of that person in the first place.

In Chapter 7 I have provided a new argument for incorporating a distinction between default and deviant values into causal models. Often the distinction is motivated by the Problem of Isomorphism. But it has been shown that instances of the Problem of Isomorphism are more straightforwardly solved by adjusting at least one of the involved causal models. But there is another problem: the Problem of Disagreement. Halpern and Hitchcock have argued that when two (or more) agents agree on the relevant variables and structural equations of a causal model but disagree with regard to causal judgements, then the default/deviant distinction is an appropriate way to reflect the disagreement. If this argument is understood as being based on descriptive claims about how agents actually reason about causes, then it does not seem to work as an argument for the default/deviant distinction. The reason is that agents who disagree about actual causes typically also disagree about which variables are to be included into the underlying causal model. However, I have suggested a prescriptive reading of this argument: agents who disagree about actual causation *should* base their claims on the same model. This helps disentangling normative and epistemic reasons for disagreement over causation. I have illustrated the relevance of this claim with an example of disagreement over causal claims regarding Search and Rescue missions in the Central Mediterranean.

Finally, let me briefly address opportunities for future work that arise from my account. First, I do not claim that the pluralist account that I have provided here is exhaustive. Looking at cases of redundancy I have argued that we need to distinguish total, path-changing, and contributing actual causation. But there may as well be further concepts of actual causation that are needed to account for other kinds of situations. Moreover, my pluralist theory has focused on actual causation as a notion that facilitates intervention and the ascription of responsibility. Again, I do not think that the resulting functional account is exhaustive. Future work may examine other purposes such as explanation and prediction, which may very well add new valuable perspectives on why we reason in terms of actual causation in the way that we do.

Another way to build on the work provided here is to apply the pluralist account as a framework for taxonomizing actual causes, for example, in epidemiology. Media coverage of the COVID-19 pandemic reports the total number of deaths that were caused by the SARS-CoV-2 virus. The relevant notion of causation that is being employed here is a notion of actual causation. But often COVID-19 patients suffer from a range of pre-existing diseases. Therefore, there arise interesting questions regarding how the virus is being selected as the salient factor and whether there are cases that involve redundancy (such that the virus is only a contributing actual cause or not a cause at all because it was preempted). Moreover, the COVID-19 outbreak can in some instances be seen as a path-changing actual cause of death. There is a significant number of cases where death can be prevented through intensive care. However, the outbreak of the COVID-19 pandemic has also lead to a breakdown of the health care systems in several countries. This has adverse consequences for patients that need intensive care for other reasons. The deaths that result from a COVID-19 induced breakdown of health care systems may, thus, also have to be counted as being actually caused by the SARS-CoV-2 virus.

Bibliography

Alicke, Mark. Culpable causation. *Journal of Personality and Social Psychology*, 63(3): 368–378, Sep 1992.

Alicke, Mark D., Rose, David, and Bloom, Dori. Causation, norm violation, and culpable control. *The Journal of Philosophy*, 108(12):670–696, 2011.

Armstrong, David. *What is a Law of Nature?* Cambridge University Press, Cambridge, UK, 1983.

Baumgartner, Michael. A regularity theoretic approch to actual causation. *Erkenntnis*, 78:85–109, 2013.

Baumgartner, Michael and Falk, Christoph. Boolean difference-making: A modern regularity theory of causation. *The British Journal for the Philosophy of Science*, forthcoming.

Baumgartner, Michael and Fenton-Glynn, Luke. Introduction to special issue on 'actual causation'. *Erkenntnis*, 78(S1):1–8, 2013.

Beale, Joseph H. The proximate consequences of an act. *Harvard Law Review*, 33(5): 633–658, 1920.

Bear, Adam and Knobe, Joshua. Normality: Part descriptive, part prescriptive. *Cognition*, 167:25–37, 2017.

Beck, Lewis White. Constructions and inferred entities. In Feigl, Herbert and Brodbeck, May, editors, *Readings in the Philosophy of Science*, pages 368–382. Appleton-Century-Crofts, New York, 1953.

Beckers, Sander and Vennekens, Joost. A general framework for defining and extending actual causation using CP-logic. *International Journal of Approximate Reasoning*, 77:105–126, 2016.

Beebee, Helen. Causing and nothingness. In Hall, Ned and Paul, L. A., editors, *Causation and Counterfactuals*, pages 291–308. MIT Press, Cambridge, MA, 2004.

Bird, Alexander. The dispositionalist concept of laws. *Foundations of Science*, 10: 353–370, 2005.

Blanchard, Thomas and Schaffer, Jonathan. Cause without default. In Helen Beebee, Huw Price, Christopher Hitchcock, editor, *Making a Difference*, pages 175–214. Oxford University Press, Oxford, 2017.

Cane, Peter. *Responsibility in Law and Morality*. Hart Publishing, Oxford, 2002.

Cappelen, Herman. *Fixing Language*. Oxford University Press, Oxford, 2018.

Carpenter, Charles E. Workable rules for determining proximate cause–part I. *California Law Review*, 20(3):229–259, 1932.

Cartwright, Nancy. Causal laws and effective strategies. *Noûs*, 13(4):419–437, 1979.

Cartwright, Nancy. *Hunting Causes and Using Them*. Cambridge University Press, Cambridge, 2007.

Chalmers, David. *The Consicus Mind. In Search of a Fundamental Theory*. Oxford University Press, Oxford, 1996.

Chockler, Hana and Halpern, Joseph. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

Collingwood, Robin George. On the so-called idea of causation. *Proceedings of the Aristotelian Society*, New Series 38:85–112, 1938.

Collins, John. Preemptive prevention. In Collins, John, Hall, Ned, and Paul, L. A., editors, *Causation and Counterfactuals*, pages 107–118. MIT Press, Cambridge, MA, 2004. Extended version of Lewis 2000.

Collins, John, Hall, Ned, and Paul, L. A., editors. *Causation and Counterfactuals*. MIT Press, Cambridge, MA, 2004.

Danks, David, Rose, David, and Machery, Edouard. Demoralizing causation. *Philosophical Studies*, 171:251–277, 2014.

Dennett, Daniel Clement. *Elbow Room*. MIT Press, Cambridge, MA, 1984.

Dowe, Phil. *Physical Causation*. Cambridge University Press, Cambridge, 2000.

Dowe, Phil. Causation and misconnections. *Philosophy of Science*, 71(5), 2004.

Ducasse, Curt John. On the nature and observability of the causal relation. *The Journal of Philosophy*, 23(3):57–68, 1926.

Edgarton, Henry W. Legal cause. *University of Pennsylvania Law Review*, 72(3): 211–244, 1924.

Eells, Ellery. *Probabilistic Causality*. Cambridge University Press, Cambridge, 1991.

Ehring, Douglas. Review of Physical Causation, Phil Dowe. *Mind*, 112(447):529–533, 2003.

Fenton-Glynn, Luke. A proposed probabilistic extension of the Halpern and Pearl definition of 'actual cause'. *The British Journal for the Philosophy of Science*, 0:1–64, 2015.

Field, Hartry. Causation in a physical world. In Loux, Michael J. and Zimmerman, Dean W., editors, *Oxford Handbook of Metaphysics*, pages 435–460. Oxgord University Press, Oxford, UK, 2003.

Fischer, David A. Causation in fact in omission cases. *Utah Law Review*, pages 1335–1384, 1992.

Fischer, David A. Insufficient causes. *Kentucky Law Review*, 94(2):277–317, 2005-2006.

Fischer, Enno. Three concepts of actual causation. *The British Journal for the Philosophy of Science*, forthcoming a.

Fischer, Enno. Causation and the problem of disagreement. *Philosophy of Science*, forthcoming b.

Frisch, Mathias. *Causal Reasoning in Physics*. Cambridge University Press, Cambridge, 2014.

Frisch, Mathias. Causation in physics. In Zalta, Edward N., editor, *The Stanford Encyclopedia of Philosophy*, forthcoming.

FRONTEX. Risk analysis for 2017, 02 2017. URL `http://frontex.europa.eu/assets/Publications/Risk_Analysis/Annual_Risk_Analysis_2017.pdf`. last accessed 31 March 2020.

Goldberger, Arthur S. Structural equation methods in the social sciences. *Econometrica*, 40(6):979–1001, 1972.

Green, Leon. Are there dependable rules of causation? *University of Pennsylvania Law Review*, 77(5):601–628, 1929.

Hall, Ned. Two concepts of causation. In Collins, John, Hall, Ned, and Paul, L. A., editors, *Causation and Counterfactuals*, pages 225–276. MIT Press, Cambridge, MA, 2004.

Hall, Ned. Structural equations and causation. *Philosophical Studies*, 132:109–136, 2007.

Halpern, Joseph Y. Defaults and normality in causal structures. In Brewka, Gerhard and Lang, Jerome, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Congress*, pages 198–108. AAAI Press, 2008.

Halpern, Joseph Y. A modification of the Halpern-Pearl definition of causality. *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 3022–3033, 2015.

Halpern, Joseph Y. *Actual Causality*. MIT Press, Cambridge, MA, 2016.

Halpern, Joseph Y. and Hitchcock, Christopher. Actual causation and the art of modeling. In Hector Dechter, Rina Geffner and Halpern, Joseph Y., editors, *Heurisitcs, Probability, and Causality: A Tribute to Judea Pearl*, pages 383–406. College Publications, London, 2010.

Halpern, Joseph Y. and Hitchcock, Christopher. Compact representations of causal models. *Cognitive Science*, 37:986–1010, 2013.

Halpern, Joseph Y. and Hitchcock, Christopher. Graded causation and defaults. *The British Journal for the Philosophy of Science*, 66:413–457, 2015.

Halpern, Joseph Y. and Pearl, Judea. Causes and explanations: a structural-model approach. Part I: Causes. *Proc. Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI2001)*, pages 194–202, 2001.

Halpern, Joseph Y. and Pearl, Judea. Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4): 843–887, 2005.

Hannikainen, Ivar and Cona, Florian. Replication of Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control., August 2017. URL `osf.io/cav4k`.

Hanson, Norwood Russell. *Patterns of Discovery*. Cambridge University Press, Cambridge, 1958.

Hart, H. L. A. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford University Press, Oxford, 1968.

Hart, H. L. A. and Honoré, A. M. *Causation in the Law*. Clarendon Press, Oxford, 1959.

Heller, Charles and Pezzani, Lorenzo. Death by rescue, 04 2016. URL `https://deathbyrescue.org/`. Last accessed 13/01/2020.

Heller, Charles and Pezzani, Lorenzo. Blaming the rescuers, 06 2017. URL `https://blamingtherescuers.org/`. Last access: 13/01/2020.

Hempel, Carl G. and Oppenheim, Paul. Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175, 1948.

Hesslow, Germund. Discussion: Two notes on the probabilistic approach to causality. *Philosophy of Science*, 43(2):290–292, 1976.

Hiddleston, Eric. Causal powers. *The British Journal for the Philosophy of Science*, 56: 27–59, 2005.

Hitchcock, Christopher. Salmon on explanatory relevance. *Philosophy of Science*, 62 (2):304–320, 1995.

Hitchcock, Christopher. The role of contrast in causal and explanatory claims. *Synthese*, 107:395–419., 1996.

Hitchcock, Christopher. The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98(6):273–299, 2001.

Hitchcock, Christopher. Of Humean Bondage. *The British Journal for the Philosophy of Science*, 54:1–25, 2003.

Hitchcock, Christopher. What Russell got right. In Corry, Richard and Price, Huw, editors, *Causation, Physics, and the Constitution of Reality. Russell's Republic Revisited*, pages 45–65. Oxford University Press, Oxford, 2007a.

Hitchcock, Christopher. Prevention, preemption, and the principle of sufficient reason. *Philosophical Review*, 66:495–532, 2007b.

Hitchcock, Christopher. How to be a causal pluralist. In Machamer, Peter K., editor, *Thinking About Causes: From Greek Philosophy to Modern Physics*, pages 200–221. University of Pittsburgh Press, Pittsburgh, PA, 2007c.

Hitchcock, Christopher. Trumping and contrastive causation. *Synthese*, 181:227–240, 2011.

Hitchcock, Christopher. Actual causation: What's the use. In Beebee, Helen, Hitchcock, Christopher, and Price, Huw, editors, *Making a Difference*, pages 116–131. Oxford University Press, 2017.

Hitchcock, Christopher and Knobe, Joshua. Cause and norm. *The Journal of Philosophy*, 106(11):587–612, 2009.

Hoerl, Christoph, McCormack, Teresa, and Beck, Sarah R. Introduction: Understanding counterfactuals and causation. In Hoerl, Christoph, McCormack, Teresa, and Beck, Sarah R., editors, *Understanding Counterfactuals, Understanding Causation*, pages 1–15. Oxford University Press, Oxford, UK, 2011.

Honoré, Tony. Necessary and sufficient conditions in tort law. In Owen, David G., editor, *The Philosophical Foundations of Tort Law*. Clarendon Press, 1997.

Hopkins, Mark and Pearl, Judea. Clarifying the usage of structural models for commonsense causal reasoning. *Proceedings of AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, pages 83–89, 2003.

Huber, Franz. Structural equations and beyond. *The Review of Symbolic Logic*, 6(4): 709–723, 2013.

Hume, David. *Enquiries Conerning Human Understanding and Concerning The Principles of Morals*. Clarendon Press, Oxford, UK, 1777/1975.

Hüttemann, Andreas. A disposition-based process theory of causation. In Mumford, Stephen and Tugby, Matthew, editors, *Metaphysics and Science*, pages 101–122. Oxford University Press, Oxford, 2013.

Hüttemann, Andreas. Processes, pre-emption and further problems. *Synthese*, pages 1–23, forthcoming.

Kahneman, Daniel and Miller, Dale. Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93:136–153, 1986.

Kim, Jaegwon. Mechanism, purpose, and explanatory exclusion. *Philosophical Perspectives*, 3:77–108., 1989.

Kirfel, Lara and Lagnado, David. "oops, Idid it again." The impact of frequent behaviour on causal judgement. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, 2017.

Knobe, Joshua and Fraser, Ben. Causal judgment and moral judgment: Two experiments. In Sinnott-Armstrong, Walter, editor, *Moral Psychology, vol. 2: The Cognitive Science of Morality: Intuition and Diversity*, pages 441–448. MIT Press, Cambridge, MA, 2008.

Kutach, Douglas. *Causation and Its Basis in Fundamental Physics*. Oxford University Press, Oxford, UK, 2013.

Lewis, David. Causation. *The Journal of Philosophy*, 70(17):556–567, 1973a.

Lewis, David. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973b.

Lewis, David. Counterfactual dependence and time's arrow. *Noûs*, 13(4):455–476, 1979.

Lewis, David. Postscripts to 'Causation'. In *Philosophical Papers Vol II*, pages 271–213. Oxford University Press, Oxford, 1986a.

Lewis, David. Causal explanation. In *Philosophical Papers Vol II*, pages 214–240. Oxford University Press, Oxford, 1986b.

Lewis, David. Events. In *Philosophical Papers Vol II*, pages 241–270. Oxford University Press, Oxford, 1986c.

Lewis, David. Causation as influence. *The Journal of Philosophy*, 97(4):182–197, 2000.

Lewis, David. Causation as influence. In Collins, John, Hall, Ned, and Paul, L. A., editors, *Causation and Counterfactuals*, pages 75–106. MIT Press, Cambridge, MA, 2004. Extended version of Lewis 2000.

Lipton, Peter. Causation outside the law. In Gross, Hyman and Harrison, Ross, editors, *Jurisprudence: Cambridge Essays*, pages 127–148. Oxford University Press, 1992.

Livengood, Jonathan. Actual causation and simple voting scenarios. *Noûs*, 47(2): 316–345, 2013.

Loewer, Barry. Why is there anything except physics? *Synthese*, 170:217–233, 2009.

Lombrozo, Tania. Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61:303–332, 2010.

Mackie, John Leslie. Causes and conditions. *American Philosophical Quarterly*, 2: 245–64, 1965.

Mackie, John Leslie. *The Cement of the Universe: A Study in Causation*. Clarendon Press, Oxford, 1974.

Malcolm, Norman. The conceivability of mechanism. *Philosophical Review*, 77:45–72., 1968.

Malone, Wex. Ruminations on cause-in-fact. *Stanford Law Review*, 9(1):60–99, 1956.

Maudlin, Tim. Causation, counterfactuals, and the third factor. In J. Collins, L. A. Paul, E. J. Hall, editor, *Causation and Counterfactuals*, pages 419–443. MIT Press, 2004.

McCain, Kevin. Intervention defended. *Logos and Episteme*, 6:61–73, 2015.

McDermott, Michael. Redundant causation. *British Journal for the Philosophy of Science*, 46:523–544, 1995.

McGrath, Sarah. Causation by omission: A dilemma. *Philosophical Studies*, 123(1): 125–148, 2005.

Menzies, Peter. Probabilistic causation and causal processes: A critique of Lewis. *Philosophy of Science*, 65(4):642–663, 1989.

Menzies, Peter. Probabilistic causation and the pre-emption problem. *Mind*, 105 (417):85–117, 1996.

Menzies, Peter. Difference-making in context. In Collins, John, Hall, Ned, and Paul, L. A., editors, *Causation and Counterfactuals*, pages 139–180. MIT Press, 2004.

Menzies, Peter. Causation in context. In Price, Huw and Corry, Richard, editors, *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*, pages 191–223. Oxford University Press, 2007.

Menzies, Peter. Platitudes and counterexamples. In Beebee, Helen, Hitchcock, Christopher, and Menzies, Peter, editors, *The Oxford Handbook of Causation*, pages 341–367. Oxford University Press, 2009.

Menzies, Peter. The problem of counterfactual isomorphs. In Beebee, Helen, Hitchcock, Christopher, and Price, Huw, editors, *Making a Difference: Essays on the Philosophy of Causation*, pages 153–174. Oxford University Press, 2017.

Menzies, Peter and Beebee, Helen. Counterfactual theories of causation. In Zalta, Edward N., editor, *The Stanford Encyclopedia of Philosophy*, Winter 2019. URL `https://plato.stanford.edu/entries/causation-counterfactual/`.

Mill, John Stuart. *A System of Logic, Rationcinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Harper & Brothers, New York, eighth edition, 1843/1882.

Moore, Michael. *Placing Blame. A Theory of Criminal Law*. Oxford University Press, Oxford, 1997.

Moore, Michael. *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*. Oxford University Press, Oxford, 2009.

Nagel, Ernest. The logic of historical analysis. In Feigl, Herbert and Brodbeck, May, editors, *Readings in the Philosophy of Science*, pages 688–700. Appleton-Century-Crofts, New York, 1953.

Nagel, Thomas. *Mortal Questions*. Cambridge University Press, Cambridge, UK, 1979.

Northcott, Robert. Causation and contrast classes. *Philosophical Studies*, 139(1): 111–123, 2008.

Pearl, Judea. On the definition of actual cause, 1998. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.9540`.

Pearl, Judea. *Causality. Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

Penzias, Arnold A. and Wilson, Robert W. A measurement of excess antenna temperature at 4080 Mc/s. *Astrophysical Journal*, 142:419–421, 1965.

Perry, Stepehn R. Loss, agency, and responsibility for outcomes: Three conceptions of corrective justice. In Feinberg, Joel and Coleman, Jules, editors, *Philosophy of Law*, pages 546–559. Wadsworth/Thompson Learning, Belmont, CA, 6th edition, 2000.

Phillips, Jonathan S. Replication of Hitchcock & Knobe (2009) Cause and Norm, September 2017. URL `osf.io/ykt7z`.

Phillips, Jonathn and Cushman, Fiery. Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, 114(18): 4649–4654, 2017.

Price, Huw and Corry, Richard, editors. *Causation, Physics and the Consitution of Reality*. Clarendon Press, Oxford, 2007.

Putnam, Hilary. Why there isn't a ready-made world. *Synthese*, 51:141–167, 1982.

Reichenbach, Hans. *The Direction of Time*. University of California Press, Berkeley, CA, 1956.

Rose, David and Danks, David. Causation: Empirical trends and future directions. *Philosophy Compass*, 7(9):643–653, 2012.

Rosenberg, Ian and Glymour, Clark. Review of Joseph Halpern, Actual Causality. *The British Journal for the Philosophy of Science*, 2018. URL `http://www.thebsps.org/2018/07/joseph-y-halpern-actual-causality/`.

Roxborough, Craig and Cumby, Jill. Folk psychological consepts: Causation. *Philosophyical Psychology*, 22(2):205–213, 2009.

Salmon, Wesley. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton, NJ, 1984.

Salmon, Wesley C. Statistical explanation. In Salmon, Wesley C., Jeffrey, Richard C., and Greeno, James G., editors, *Statistical Explanation and Statistical Relevance*, pages 29–88. University of Pittsburgh Press, Pittsburgh, PA, 1971.

Samland, Jana and Waldmann, Michael. How prescriptive norms influence causal inferences. *Cognition*, 156:164–176, 2016.

Schaffer, Jonathan. Trumping preemption. *The Journal of Philosophy*, 97(4):165–181, 2000a.

Schaffer, Jonathan. Causation by disconnection. *Philosophy of Science*, 67:285–300, 2000b.

Schaffer, Jonathan. Overdetermining causes. *Philosophical Studies*, 114:23–45, 2003.

Schaffer, Jonathan. Causes need not be physically connected to their effects. In Hitchcock, Christopher, editor, *Contemporary Debates in Philosophy of Science*, pages 197–216. Basil Blackwell, Oxford, 2004.

Schaffer, Jonathan. Contrastive causation. *The Philosophical Review*, 114(3):297–328, 2005.

Schaffer, Jonathan. Causal contextualism. In Blaauw, Martijn, editor, *Contrasitvism in Philosophy*, pages 35–63. Routledge, Oxford, 2012.

Schlick, Moritz. *Problem of Ethics*. Prentice -Hall, Inc., New York, 1939. Authorized translation by David Rynin.

Skyrms, Brian. *Causal Necessity*. Yale University Press, New Haven and London, 1980.

Smart, J. J. C. Free-will, praise and blame. *Mind*, 70(279):291–306, 1961.

Smith, Jeremiah. Legal cause in actions of tort. *Harvard Law Review*, 25(4):303–327, 1912.

Spirtes, Peter, Glymour, Clark, and Scheines, Richard. *Causation, Prediction, and Search*. Springer-Verlag, 1993.

Stalnaker, Robert. A theory of conditionals. In *Studies in Logical Theory*. Blackwell, Oxford, UK, 1968.

Stapleton, Jane. Choosing what we mean by 'causation' in the law. *Missouri Law Review*, 73(2):433–380, 2008.

Statham, Georgie. Woodward and variable relativity. *Philosophical Studies*, 175(4): 885–902, 2017.

Steel, Daniel. Cartwright on causality: Methods, metaphysics and modularity - Hunting Causes and Using Them: Approaches in Philosophy and Economics, Nancy Cartwright. *Economics and Philosophy*, 26(1):77–86, 2010.

Strawson, Peter. Freedom and resentment. In Watson, Gary, editor, *Proceedings of the British Academy*, volume 48, pages 1–25. Oxford University Press, Oxford, 1962.

Strevens, Michael. Review of Woodward, Making things happen. *Philosophy and Phenomenological Research*, 74:233–249, 2007.

Strevens, Michael. Comments on Woodward, Making things happen. *Philosophy and Phenomenological Research*, 77:171–192, 2008.

Strevens, Michael. *Depth. An Account of Scientific Explanation*. Harvard University Press, Cambridge, MA, 2011.

Suppes, Patrick. *A Probabilistic Theory of Causality*. North-Holland Publishing Company, Amsterdam, 1970.

Sytsma, Justin, Livengood, Jonathan, and Rose, David. Two types of typicality: Rethinking th role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43:814–820, 2012.

Tooley, Michael. The nature of laws. *Canadian Journal of Philosophy*, 7(4):667–698, 1977.

Tversky, Amos and Kahneman, Daniel. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5:207–232, 1973.

Tversky, Amos and Kahneman, Daniel. Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293–315, 1983.

van Fraassen, Bas C. *The Scientic Image*. Oxford University Press, 1980.

Vargas, Manuel. Moral influence, moral responsibility. In Trakakis, Nick and Cohen, Daniel, editors, *Free Will and Moral Responsibility*, pages 90–122. Cambridge Scholars Press, Newcastle, 2008.

Vincent, Nicole A. A structured taxonomy of responsibility concepts. In Vincent, Nicole A, van de Poel, Ibo, and van den Hoven, Jeroen, editors, *Moral Responsibility. Beyond Free Will and Determinism*, pages 15–36. Springer, 2011.

Walker, Ian. Drivers overtaking bicyclists: Objective data on the effects of riding position, helmet use, vehicle type and apparent gender. *Accident Analysis and Prevention*, 39:417–425, 2007.

Walsh, Clare R. and Sloman, Steven A. The meaning of cause and prevent: The role of causal mechanism. In Bara, Bruno G., Barsalou, Lawrence, and Bucciarelli, Monica, editors, *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 2331–2336, Mahwah, NJ, 2005. Lawrence Erlbaum Associates.

Walzer, Michael. *Thick and Thin: Moral Argument at Home and Abroad*. Universtiy of Notre Dame Press, Notre Dame, IN, 1994.

Waters, C. Kenneth. Causes that make a difference. *Journla of Philosophy*, 104(11), 2007.

Williams, Bernard. *Moral Luck*. Cambridge University Press, Cambridge, UK, 1981.

Wilson, Robert W. The cosmic microwave background radiation. Nobel Lecture, 8 December, 1978. URL `http://www.nobelprize.org/nobel_prizes/physics/laureates/1978/wilson-lecture.pdf`.

Woodward, James. *Making Things Happen*. Oxford University Press, Oxford, 2003.

Woodward, James. Response to Strevens. *Philosophy and Phenomenological Research*, 77:193–212, 2008.

Woodward, James. Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology and Philosophy*, 25(3):287–318, 2010.

Woodward, James. 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, 2014.

Woodward, James. Methodology, ontology, and interventionism. *Synthese*, 192: 3577–3599, 2015.

Woodward, James. The problem of variable choice. *Synthese*, 193:1047–1072, 2016.

Wright, Richard. Causation in tort law. *California Law Review*, 73(6):1735–1828, 1985.

Wright, Richard. Once more into the bramble bush: Duty, causal contribution, and the extent of legal responsibility. *Vanderbilt Law Review*, 54(3):1–51, 2001.

Wright, Richard. The NESS account of natural causation: A response to criticisms. In Goldberg, R., editor, *Perspectives on Causation*, pages 265–322. Hart Publishing, Oxford, 2011.

Wright, Sewall. The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences of the United States of America*, 6(6):320–332, 1920.

Wright, Sewall. Correlation and causation. *Journal of Agricultural Research*, 20(7): 557–585, 1921.