Causal Explanations
How to Generate, Identify, and Evaluate Them

A Causal Model Approach Augmented with Causal Powers

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Philosophie
der Ludwig-Maximilians-Universität
München

vorgelegt von
Jan Borner

aus
Waldbröl

2023

Referent/in:	Prof. DDr. Hannes Leitgeb
Korreferent/in:	Prof. Dr. Christian List

Tag der mündlichen Prüfung: 13. Februar 2023

Acknowledgements

I would like to thank those who accompanied and supported me during the time of my PhD project.

First and foremost, I want to thank Hannes Leitgeb. Early on in my studies, his work and his teaching have made me realize how valuable and fruitful formal methods can be within philosophy and he inspired me to pursue this kind of philosophy further and further. As the supervisor of my PhD project he always motivated me to pursue exactly those questions that mattered to me. In so many conversations his advice and suggestions has helped me immensely in improving my work.

I also want to thank Christian List and Stephan Hartmann for invaluable feedback on still way too long drafts. I have benefited enormously from their assessments, tips and comments.

I am grateful for many great talks and events at the MCMP and I have received a lot of helpful and constructive feedback from the participants of Hannes Leitgeb's graduate seminars and Christian List's research seminar. I would also like to thank Borut Trpin and Matías Osta-Vélez for organizing the reading group on reasoning and inference, which has produced many exciting discussions. Borut Trpin, Rolf Pfister, and Fabian Fischer also provided very helpful comments on some early drafts.

The *Studienstiftung des Deutschen Volkes* has generously sponsored my PhD project, for which I am enormously grateful.

I am deeply thankful to Fabian Fischer and Sebastian Hengst for so many great discussions and conversations not only about philosophy, but also about the process of writing a dissertation itself.

My friends and my parents have accompanied and supported me through all the ups and downs during my PhD project. I would especially like to thank my godmother, Marita Schmalenbach, who encouraged me from an early age to follow my path and who unfortunately did not live to see the completion of my dissertation.

And then there is Roschelle, who has accompanied me as my girlfriend, my fiancée, and finally as my wife during the years of my PhD project. Her support, her love, and her patience with me has been enormous and she has been the causal explanation for all the best moments over the last couple of years.

Statement on Prior Publications

Chapter 1 is an adapted and expanded version of an article of mine that has already been accepted for publication by *The British Journal for the Philosophy of Science*:

Jan Borner. *Halpern and Pearl's Definition of Explanation Amended.* Accepted for publication by *The British Journal for the Philosophy of Science* on August 08, 2021.
https://doi.org/10.1086/716768. Copyright 2021 The British Society for the Philosophy of Science.

I want to thank two anonymous referees for *The British Journal for the Philosophy of Science* for many valuable comments.

Contents

List of Figures

List of Tables

Introduction

Recently, I received a phone call on my landline where normally barely anyone calls. The voice on the other end of the line sounded strange, somewhat mechanical, and it started to speak with a weird delay after my curious “Hello?”. “This is a message from Interpol” it said. It told me about an identity theft, due to which my cooperation was now urgently needed. Refusing to cooperate could have severe consequences. I did, what probably most people would do in a situation like this. I hung up the phone. That Interpol would call me and ask for my cooperation after an identity theft by using a badly recorded tape was just too weird to be true. Instead, I concluded that scammers were behind the tape, simply calling people at random in the hope that at least a few would believe their story and be willing to send them money or personal information. I was so sure of my conclusion that I accepted the risk of the “severe consequences” that I would face, if I were wrong.

And yet, my conclusion was not the result of a deductive argument. Given my knowledge at that time, I could not have used truth-preserving rules of inference to reach my conclusion. I also did not reach it by an inductive argument that is based on repeated observations or even a statistically representative data set. It was the first and only time I received such a call and at that time I have never heard of it before. I was nevertheless sure of my conclusion. Even if I assumed, that I had actually been the victim of an identity theft, it would not really explain, why Interpol was directly reaching out to me and asking for my cooperation, especially by playing a tape recording without even caring to know if it was really me who picked up the phone. But the assumption that scammers were behind the call, was able to explain everything that was so strange about it: The anonymity of the recorded message, the lack of concern about who was actually receiving the message, and the ominous warning of severe consequences. My conclusion was the result of an Inference to the Best Explanation.

Inference to the Best Explanation (IBE) is a form of reasoning that is omnipresent in mundane as well as in delicate situations with serious and far-reaching consequences. An Inference to the Best Explanation is not only helpful when it comes to recognizing a scam. It helps the detective solve a case. It helps the judge decide on the guilt of a defendant. It helps the doctor diagnose. And it helps the scientist evaluate hypotheses. And yet, it is barely understood. While formal accounts of deductive and inductive reasoning have been flourishing for quite some while, IBE has long lived in the shadows in the field of formal philosophy.

But that is changing now. After authors like Harman (1965) and Lipton (2004) pioneered and pointed to the importance of Inference to the Best Explanation as a legitimate form of ampliative reasoning, IBE is receiving increasing attention.[1] There are by now several proposals on how

[1] A very early pioneer for IBE is, of course, Charles Sanders Peirce. See, for example, (Gabbay and Woods,

IBE can be formally explicated in precise terms. For example, Boutilier and Becher (1995) as well as Pagnucco (1996) try to formalize IBE by employing the framework of Alchourrón, Gärdernfors, and Makinson's (AGM) belief revision theory.[2] Aliseda (1997) uses the proof procedure of semantic tableaux to formulate an algorithm that, for a given background theory, produces explanation candidates for a given event to be explained (explanandum). Another algorithm with exactly the same purpose is proposed by Meheus and Provijn (2007), who build on Batens and Provijn (2003) and their formalism of goal-directed proofs, a decision method for testing whether a given formula is derivable from a given set of premises. Using the framework of adaptive logics,[3] Meheus and Batens (2006) put forward yet another formalization of IBE that is later advanced by Meheus (2011).

All these accounts opened up promising paths that may ultimately lead to suitable and precise formal models of IBE. However, in their current form they also suffer from the same significant weakness. They all embrace and build upon an outdated and highly simplistic picture of explanation, which essentially corresponds to Hempel and Oppenheim's (1948) deductive-nomological (DN) account of explanation.[4] The DN-account basically says that a hypothesis is able to explain a given explanandum if the explanandum can be derived from the hypothesis (and some additional background assumptions) by a deductively valid argument. Even though the DN-account had for a long time been the received and largely unchallenged model of explanation, it has meanwhile clearly lost this status. The philosophical literature has by now revealed several weaknesses of the DN-account and has produced plenty of counterexamples, in which the DN-model identifies something as an explanation that intuitively is clearly none.[5] As a consequence, none of the formal accounts of IBE mentioned above is able to ensure that the hypothesis, which is inferred, really amounts to what we intuitively perceive as an explanation.

In my dissertation, I will not provide a complete formalization of IBE. At this point, it would be far too comprehensive a project to complete in the context of a single dissertation. Instead, I will focus on laying a solid foundation that is needed for an adequate formalization of IBE. This foundation will consist of three major components. First, an intuitively adequate and formally precise model of explanation. Second, an intuitively adequate and formally precise measure of explanatory power. And third, an intuitively adequate and formally precise criterion of proportionality that is able to identify the most appropriate level of specificity of a given explanation.

Part I: Causal Explanations

The main motivation for the first part of my dissertation is based on the conviction that an adequate formal account of Inference to the Best Explanation needs an adequate formal account of explanation. It is the goal of the first part to provide such an account of explanation.

Ever since the DN-account was pushed off the throne, there has been a decades-long struggle in the philosophical literature to find an appropriate and worthy successor as a commonly-

2005, p. 75 ff.) for a concise exposition of Peirce's views on IBE and abductive reasoning.

[2] See (Gärdenfors, 1988) for a comprehensive presentation of the AGM belief revision theory.

[3] See (Straßer, 2014) for a detailed introduction to adaptive logics.

[4] See also (Hempel, 1965) for an in depth exposition of the DN-account.

[5] See, for example, (Salmon, 1989, pp. 46-50) for a list of famous counterexamples to the DN-account of explanation.

received model of explanation. Up until now, no single model has been able to establish itself as an universally accepted and all-encompassing account of explanation. Nonetheless, it has become something like a new consensus that a causal conception of explanation appears to be particularly promising, because it has the potential to successfully deal with two major problems that many other accounts of explanation struggle with. A causal approach is able to reliably differentiate between explanatory relevant and irrelevant information for a given explanandum by equating explanatory relevance with causal relevance. And a causal approach is able to account for the asymmetry of explanations due to the asymmetry of causation. Accordingly, several accounts of causal explanation have been proposed over the last few decades.[6]

I will not explicitly discuss and compare all the competing accounts of explanation that have been put forward so far. Other authors have already done this in detail and so well and extensively that I could barely add anything essential to it.[7] Instead, this dissertation will take an account as its point of departure that I consider to be one of the most adequate and formally precise models of explanation to date, namely the account of causal explanation developed by Halpern and Pearl (2005b). Interestingly, the Halpern and Pearl account (HP-account) of explanation, which Halpern (2016) has slightly revised and updated, seems to be widely unknown within philosophy. After its publication in (Halpern and Pearl, 2005b), there have barely been any reactions to it in the philosophical literature. But for the project at hand, the HP-account of explanation turns out to be something like a hidden gem. There are three main reasons for this.

First, the HP-account of causal explanation is built upon one of the most established and elaborate formal accounts of causation, namely a causal-model-based, interventionist account of causation.[8] The embedding in this formal framework enables Halpern and Pearl to provide a precise formal definition of causal explanation.

Secondly, the HP-account aims to explicate a concept of *potential* causal explanation, that encompasses facts or events for which it is not yet certain whether they actually explain a given phenomenon. Instead, they only have a certain possibility of doing so. This includes, in particular, facts that are not yet known to be true. This is exactly what we need for a formalization of IBE, since the hypotheses that are weighed up in an Inference to the Best Explanation are not yet known to be true. Otherwise, an Inference to the Best Explanation would not be necessary in the first place.

Finally, the HP-account considers a causal explanation to be relative to the epistemic state of a given agent. Thus, according to the HP-account, any causal explanation of a given phenomenon is always an explanation of that phenomenon *for* a certain agent in a certain epistemic state. This again is crucial for an analysis of IBE, since any Inference to the Best Explanation occurs in the context of certain background beliefs, that is, an epistemic state. The background beliefs are crucial for identifying explanation candidates of a given explanandum and for evaluating and weighing up the identified explanation candidates to select the best of them.[9] This is why

[6] See, for example, (Salmon, 1984), (Lewis, 1986c), (Woodward, 2003), (Craver, 2007), or (Strevens, 2008).

[7] See, for example, (Salmon, 1989), (Klärner, 2003), (Woodward and Ross, 2021), or (Strevens, 2008, Part I).

[8] See (Pearl, 2000), (Woodward, 2003), or (Halpern, 2016) for thorough expositions of interventionist accounts of causation.

[9] See, for example, (Pagnucco, 1996), (Aliseda, 1997), or (Meheus and Provijn, 2007) for the importance of background beliefs or background theories for IBE.

any account of explanation that aims to provide a concept of explanation that can serve as a basis for IBE, needs to acknowledge the epistemically relative nature of potential explanations. This is exactly what the HP-account does and it thereby stands out from other, more well-known accounts of causal explanation, like the ones developed by Salmon (1984), Lewis (1986c), Woodward (2003), Craver (2007), or Strevens (2008).

The three aspects just mentioned make Halpern and Pearl's account of causal explanation a very promising candidate for buidling a formalization of IBE on it. Nonetheless, after a short introduction into the framework of causal models, I will argue in chapter 1 that the HP-account of explanation still struggles with several inadequacies. Based on this criticism, I develop a new account of causal explanation that retains the aforementioned positive features of the HP-model, but also overcomes its demonstrated shortcomings. The resulting account will consist of several definitions of distinct, but related concepts of causal explanation, including the concepts of *potential*, *actual*, and *partial explanation*.

In chapter 1, I follow Halpern and Pearl (2005b) and Halpern (2016) in making a crucial simplification: I only consider contexts with deterministic causal relationships. But very often, we can only describe causal relationships probabilistically. In chapter 2, I will therefore explore how probabilistic causal relationships can be introduced into the framework of causal models. For this purpose, I am going to compare two related, but slightly different formal methods. The first method employs Causal Bayes Nets (CBNs), while the second uses Structural Equation Models (SEMs) with error-terms. I will argue that the two approaches ultimately lead to two different interpretations of token causation. Based on the SEM-approach, I will show in chapter 3 how the definitions of explanation, as put forward in chapter 1, can be applied in contexts with probabilistic causal relationships. This concludes the first part of my dissertation.

Part II: Causal Explanatory Power

For a formalization of IBE, it is not enough to have a formal account of potential explanation, though. After all, IBE is all about inferring the best from a selection of potential explanations. For this it needs a measure to compare and weigh up different explanation candidates. It is the second major goal of this dissertation to explicate such a measure. In chapters 4 to 8, I will work towards this goal.

I will start in chapter 4 by exploring the Power PC theory that has been developed by the psychologist Patricia Cheng (1997). The theory drafts a measure of what is supposed to be the intrinsic generative causal power of a cause on an effect. I will defend Cheng's measure against recent criticism and I will expand Cheng's Power PC theory to make her measure applicable to more complex causal scenarios than it was originally designed for. In chapter 5, I will do the same with a measure that is supposed to quantify the adversary of intrinsic generative causal power, namely the intrinsic preventive power of a preventive cause.

In the process of explicating the measures of generative causal power and preventive power, we will frequently encounter two concepts of token causation that play a crucial role in Cheng's Power PC theory, but that do not appear in standard causal-model-based interventionist accounts of causation, namely the concepts of causal production and prevention. While the expanded Power PC theory, as developed in chapters 4 and 5, already describes several core features of both concepts, I aim for a more thorough analysis in chapter 6. Specifically, I am going to explore how the concepts of causal production and prevention correspond to interventionist concepts of token causation, which form the foundation for the account of causal explanation developed in the first three chapters. This exploration ultimately motivates the definition of yet another concept of causal explanation, which I am going to put forward in chapter 7 under the label of *extensive causal explanation*. The idea underlying this concept is that an extensive causal explanation explicitly contains all the information that is necessary to evaluate its power. I will argue that any extensive causal explanation conforms to the schema 'ϕ, *because* ψ, *despite* χ' and I will put forward an algorithm for generating extensive causal explanations in probabilistic causal models.

In chapter 8, I will then discuss three prominent proposals for measuring a form of explanatory goodness, that has come to be known as *explanatory power*: The measure by Schupbach and Sprenger (2011), the measure by Crupi and Tentori (2012), and the measure by Good (1960) and McGrew (2003). I will argue that all three measures are inadequate as measures of explanatory power, because all three measures often produce highly unintuitive results. I will therefore put forward a new method for determining the explanatory power of a causal explanation. This method employs the measures of intrinsic generative causal power and intrinsic preventive power as developed in chapters 4 and 5. This concludes the second part of my dissertation.

Part III: Proportional Causal Explanations

In the third part of my dissertation, I will turn to another concept that is sometimes discussed as a potential criterion of explanatory goodness when it comes to causal explanations, namely the concept of *proportionality*.[10] The idea underlying this concept, as it is introduced by Yablo (1992), is that a cause is proportional relative to a given explanandum, if it is neither too abstract, nor too specific in respect to the explanandum. But proportionality is only reasonable as a criterion of explanatory goodness, if causal claims on different levels of abstraction, that is, on different levels of supervenience, are possible in the first place.

In chapter 9, I will therefore deal with Kim's (2005) argument of causal exclusion and I will argue that it does not apply to the concepts of causation underlying my account of causal explanation. I will then address the problem that in the classical framework of causal models, it is impossible to express and evaluate causal claims on different supervenience levels relative to a single causal model. For dealing with this problem, I will propose an amendment to the classical causal model framework by incorporating a formalism developed by List (2019), namely Systems of Levels.

In chapter 10, I will employ the newly amended causal model framework to formulate a new definition of proportionality. I will argue that this concept of proportionality is indeed a valuable feature for any potential causal explanation that serves as a candidate in an Inference to the Best Explanation. This will conclude the third and final part of my dissertation.

The accounts of causal explanation, causal explanatory power, and proportionality, that I am going to develop in the following ten chapters, are supposed to adhere to the standards of

[10] See, for example, (Woodward, 2010), (Woodward, 2018), or (Kinney, 2019a).

a Carnapian explication.[11] They should correspond as far as possible to what we intuitively understand by these terms (similarity). The accounts should be formally precise (exactness). They should be fruitful for the project at hand, which is ultimately a formalization of IBE (fruitfulness). And they should avoid unnecessary complexity (simplicity).

While I hope that the results of the following ten chapters will put us in a much better position to analyze and ultimately formalize a method of IBE, I also think that they are valuable in their own right. They will not only provide us with a better grasp of the workings of explanations, but they will also enhance our understanding of the stuff that Mackie (1974) has called "the cement of the universe", namely causation. In the following, I hope to give an insight into how this cement is mixed.

[11] See, for example, (Carnap, 1950), where Carnap outlines the method of explication. For some recent perspectives on explication and a depiction of the method's historical development see, for example, (Brun, 2016) and (Leitgeb and Carus, 2022).

Part I

Causal Explanations

Chapter 1

Causal Explanations in Deterministic Contexts

1.1 Introduction

In the second part of their two-part article *Causes and Explanations: A Structural-Model Approach* (2005b) Halpern and Pearl use the framework of structural equation models to define a notion of explanation that is based on the concept of actual causation. Besides its formal clarity the definition has several promising features: it adequately captures common intuitions about causal explanations in many different scenarios, it accounts for pragmatic features of explanations by taking into account the epistemic state of the explanation's recipient, it can account for contrastive explanations,[1] the approach can be expanded to cover scenarios with probabilistic causal relationships,[2] and it provides a fruitful framework for the formal definition of several measures of explanatory goodness. These are good enough reasons to take note of Halpern and Pearl's definition (HP-definition) of explanation. But its impact on the philosophical literature has been rather meager so far. While Halpern and Pearl's definition of actual causation as presented in (Halpern and Pearl, 2005a) has unleashed a vibrant discussion full of constructive criticism,[3] the authors' approach to explanation in (Halpern and Pearl, 2005b) has so far not been subject to the same levels of scrutiny.

The first chapter of this dissertation is here to change that. I aim to show that, even if we presuppose adequately defined concepts of causation, the HP-definition of explanation still faces serious problems and is in need of adjustments.

After giving a short introduction to the framework of structural equation models (section 1.2), I will present the HP-definition of explanation in section 1.3 as it is put forward by Halpern (2016), who reformulated the definition from (Halpern and Pearl, 2005b) by building it on a more intuitive notion of sufficient causation. I will then introduce several examples that turn out to

[1] A contrastive explanation is an explanation that not only explains why a certain event happened but why a certain event rather than some alternative happened. Halpern and Pearl (2005a, p. 859) and Fenton-Glynn (2017, p. 1080) show how contrastive causation can be explicitly represented by Halpern and Pearl's approach. This can serve as a foundation for representing contrastive causal explanations as well.

[2] This will be the topic of chapter 2 of this dissertation.

[3] See, for example, (Hall, 2007), (Halpern and Hitchcock, 2015), (Halpern, 2015), (Beckers and Vennekens, 2018), (Weslake, forthcoming), (Andreas and Günther, 2021), or (Beckers, 2021).

be problematic for the HP-definition of explanation (section 1.4). In section 1.5, I will propose several changes to the HP-definition to yield a definition of explanation that is able to handle the given examples successfully. I will then deal with partial explanations in section 1.6 and with a measure that is able to quantify the degree of explanatory completeness of partial explanations. Finally, in section 1.7, I will discuss an argument by Craver (2007) that threatens to undermine the virtue of the HP-approach to explanation, including my amendments of the HP-definition, entirely. I will show that the HP-approach is actually reconcilable with Craver's position and that it is able to formally explicate conceptions of explanation that most accounts of causal explanation have neglected.

1.2 Structural Equation Models

Since the HP-definition of explanation is formulated in the framework of structural equation models, I will start with a quick introduction to the formal framework.[4]

In a *structural equation model* (or simply: *causal model*) events are represented in terms of random variables that can take on different values. Relations of direct causal influence between events are represented by structural equations which in turn encode counterfactual dependences between events. Consider a simple example from (Halpern and Pearl, 2005a), in which we aim to represent the causal structure of the following situation: two events might cause a forest fire. Either there is an arsonist, who drops a lit match in the forest, or there is a lightning strike, which ignites the fire. To model this situation we introduce a binary variable F, which takes on the value 1 ($F = 1$) if the fire actually takes place and the value 0 ($F = 0$) if it does not. Further, we introduce the binary variable L with $L = 1$ representing the fact that a lightning strike takes place and $L = 0$ representing that it does not. We also introduce the variable A with $A = 1$ representing that an arsonist drops a match in the forest and $A = 0$ meaning that no one is dropping a match in the forest. The causal influence of the arsonist and the lightning strike on the forest fire is represented by the structural equation $F := A \vee L$, which tells us that F takes on the value 1 if either $A = 1$ or $L = 1$.[5] This means that both $A = 1$ and $L = 1$ are sufficient to single-handedly produce $F = 1$. The causal structure of the situation can also be represented by a causal diagram, in which the directed edges between the variables represent relations of direct causal influence:

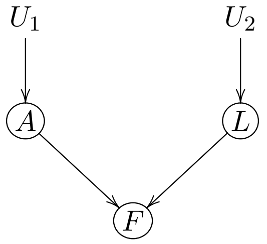


Figure 1.1: Causal diagram of the forest fire example.

[4] For a thorough discussion of structural equation models, see (Pearl, 2000).

[5] $F := A \vee L$ stands for $F := max(A, L)$. This notation makes sense only for binary variables, in which case the variable can be interpreted as a proposition and the values 0 and 1 as truth values.

In addition to the so called *endogenous variables* A, L, and F there are two *exogenous variables* U_1 and U_2. In a way, the exogenous variables lie on the representational frontier of a causal model $\mathcal{M}$. While $\mathcal{M}$ represents how the values of its endogenous variables causally depend on the values of other variables in $\mathcal{M}$ in terms of structural equations, it does not analyze how the values of its exogenous variables come about. By providing a finite number of starting points for the representation of a causally connected sequence of events the use of exogenous variables makes it possible to represent the causal structure of just a very small snippet of the overwhelmingly complex world. To serve as representational starting points, exogenous variables subsume several factors outside of the model that causally influence the value of an endogenous variable. U_1, for example, takes on the value 1 if and only if there is some event that causally produces $A = 1$. So, the causal relationship between U_1 and A is given by the structural equation $A := U_1$. Analogously, the causal relationship between U_2 and L is given by $L := U_2$.

We are now in a position to summarize the basic components of a causal model in a more formal definition.[6] A causal model $\mathcal{M}$ is a pair $(\mathcal{S}, \mathcal{F})$. $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$ is a *signature* with $\mathcal{U}$ being a set of exogenous variables, $\mathcal{V}$ being a set of endogenous variables and $\mathcal{R}$ being a function that assigns to every $Y \in \mathcal{U} \cup \mathcal{V}$ a range of values $\mathcal{R}(Y)$. $\mathcal{F}$ is a set of structural equations with exactly one structural equation F_X for each $X \in \mathcal{V}$. If F_X is a structural equation with variable X on the left-hand side and variable Y appears on the right-hand side of F_X, then we will say that Y is a *parent* of X and that X is a *child* of Y.

Another crucial ingredient in the framework of causal models is the intervention operator. An intervention $do(X = x)$ is an external manipulation that changes or fixes the value of the variable X to x regardless of the values of those variables in $\mathcal{V}$ that have a causal influence on X and without changing the value of any other variable in $\mathcal{V}$ unless those changes are causally mediated by X.[7] Every structural equation in a causal model is considered to be invariant under interventions on the variables on the right-hand side of the equation. This means that the equation correctly describes how the value of the variable on the left-hand side would change if the variables on the right-hand side would be changed by interventions. But a structural equation is not invariant under interventions on the variable on the left-hand side. Since an intervention on the variable X determines its value regardless of the values of those variables in $\mathcal{V}$ that have a causal influence on X, an intervention on X effectively disrupts the causal relations that are represented by the structural equation F_X. An intervention $do(X = x)$ can therefore be formalized as an operation on a causal model $\mathcal{M}$ that yields a new causal model $\mathcal{M}_{do(X=x)}$, which only differs from $\mathcal{M}$ insofar that the structural equation F_X is replaced by the assignment $X = x$.

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, we can define a formal language that consists of:[8]

- *primitive events*, which are formulas of the form $X = x$ with $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. (I will use the variables ϕ and ψ to represent primitive events or Boolean combinations thereof);

- *(intervention) counterfactuals*, which are formulas of the form $[Y_1 \leftarrow y_1, ..., Y_k \leftarrow y_k]\phi$ (read as: If Y_1 would be set to y_1 by intervention and ... and Y_k would be set to y_k by intervention,

[6] I follow the formal exposition of causal models as presented by Halpern and Pearl (2005a).

[7] See (Woodward, 2003, p. 98) for a more detailed characterization of interventions.

[8] Here again, I follow the notation introduced by (Halpern and Pearl, 2005a, p. 852)

then ϕ) with $Y_1, ..., Y_k$ being distinct variables in $\mathcal{V}$ and $y_1 \in \mathcal{R}(Y_1), ..., y_k \in \mathcal{R}(Y_k)$.

In a recursive causal model, that is a causal model that can be represented by a directed acyclic graph, the values of all endogenous variables are completely determined by the values of the exogenous variables. An assignment $\vec{u}$ of the exogenous variables $\vec{U}$ is called a *context* and the pair $(\mathcal{M}, \vec{u})$ of a causal model and a context is a *causal setting.*[9] $(\mathcal{M}, \vec{u}) \models X = x$, in words: $X = x$ is true in $(\mathcal{M}, \vec{u})$, if the variable X has the value x in the causal setting $(\mathcal{M}, \vec{u})$. For Boolean combinations ϕ of primitive events $(\mathcal{M}, \vec{u}) \models \phi$ is defined in the usual way. Finally, $(\mathcal{M}, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\phi$ if and only if $(\mathcal{M}_{do(\vec{Y}=\vec{y})}, \vec{u}) \models \phi$.

1.3 From Causes to Explanations

1.3.1 Actual and Sufficient Causes

Halpern and Pearl build their concept of explanation on causation in the sense that an explanation 'if found to be true, would constitute a genuine cause of the explanandum' (Halpern and Pearl, 2005b, p. 897). So, to get to the definition of explanation, we first have to clarify what it means to be a genuine cause. The concept of *actual causation*, sometimes called *token causation*, is a straightforward candidate when it comes to causal explanations of token events. There is an ongoing debate about the best definition of actual causation. Halpern (2016) has put forward several definitions himself. A discussion of all different proposals is beyond the scope of this dissertation and it is not really needed for our purposes. While the explication of causal explanation that we will discuss does indeed presuppose and build on a concept of actual causation, the specific explication of this concept will be replaceable in our discussion, as long as the explication is given in the framework of causal models and as long as it conforms to the basic interventionist leitmotif that actual causation is a form of de facto dependence.[10] I will therefore base the following discussion on the definition of actual causation as given by Halpern and Pearl (2005a, p. 853), for the simple reason that it is one of the most prominent definitions in the literature and that it deals adequately with all the examples discussed in this dissertation. But whenever I refer to actual causation in the following discussion, the reader may simply plug in his or her favorite interventionist explication of this concept. Let me now first present Halpern and Pearl's complete definition of actual causation, before I discuss its basic idea:[11]

Actual Cause (Halpern and Pearl, 2005a). $\vec{X} = \vec{x}$ *is an 'actual cause' of* ϕ *in the causal setting* $(\mathcal{M}, \vec{u})$ *(in short:* $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{x} \rightsquigarrow \phi$*) if and only if the following conditions hold:*

AC1 $(\mathcal{M}, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \phi$.

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

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

[9] Like Halpern and Pearl (2005a), I write '$\vec{U}$' for a sequence $(U_1, ...U_k)$ of variables and '$\vec{u}$' for the corresponding values $(u_1, ...u_k)$ with $u_1 \in \mathcal{R}(U_1), ..., u_k \in \mathcal{R}(U_k)$. Similarly, I use '$\vec{X} = \vec{x}$' for a conjunction of primitive events $X_1 = x_1 \wedge ... \wedge X_k = x_k$.

[10] We will shortly see what is meant by 'de facto dependence'.

[11] For a deeper discussion of the concept of actual causation, see (Halpern and Pearl, 2005a) or (Halpern, 2016).

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

AC3 $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\vec{X}' = \vec{x}'$ *(with* $\vec{x}'$ *being the restriction of* $\vec{x}$ *to the variables in* $\vec{X}'$*) satisfies conditions AC1 and AC2.*

First of all, as becomes clear in the definition, the characteristic of being an actual cause of an event is always relative to a specific causal setting. There is therefore no such thing as an actual cause per se. Now, condition AC1 says that for $\vec{X} = \vec{x}$ to be an actual cause of ϕ in $(\mathcal{M},\vec{u})$, both $\vec{X} = \vec{x}$ and ϕ must be true in $(\mathcal{M},\vec{u})$. AC2(a) basically says that an effect must be counterfactually dependent on its cause, in the sense that if the cause would not have happened, then the effect would not have happened either. Most accounts of actual causation agree that the counterfactual dependence of ϕ on $\vec{X} = \vec{x}$ is a sufficient condition for $\vec{X} = \vec{x}$ being an actual cause of ϕ. But as cases of overdetermination and preemption show, it is not a necessary condition. Consider an example from Hall (2004) about Suzy and Billy, who both throw a stone at a bottle that shatters as soon as it is hit. Suzy throws a bit harder, so her stone gets to the bottle first and shatters it. But if Suzy had not thrown her stone, Billy's stone would have hit and shattered the bottle. Clearly, Suzy's throw is the actual cause of the bottle shattering, while Billy's throw is not. But because of Billy's throw the bottle shattering does not counterfactually depend on Suzy's throw. This is why in AC2(a) the condition of counterfactual dependence is weakened. AC2(a) does not demand that the effect ϕ is counterfactually dependent on its actual cause $\vec{X} = \vec{x}$ in the given causal setting $(\mathcal{M},\vec{u})$. Instead, the counterfactual dependence of ϕ on $\vec{X} = \vec{x}$ must only manifest itself in some counterfactual situation or contingency that is created by changing the actual causal setting through interventions on certain variables $\vec{W}$.[12] This is what Yablo (2002) calls a *de facto dependence*, which means that the effect counterfactually depends on the cause "with the right things held fixed" (Yablo, 2002, p. 130). The shattering of the bottle, for example, does not counterfactually depend on Suzy's throw in the actual situation, but it de facto depends on Suzy's throw, since it counterfactually depends on Suzy's throw in the contingency in which Billy is stopped from throwing through an intervention. The role of AC2(b) is, basically, to restrict the set of contingencies that are admissible for attesting a de facto dependence of ϕ on $\vec{X} = \vec{x}$, since it adds the additional condition that in the considered contingency, as well as in all contingencies lying 'in between' the actual situation and the considered contingency, setting $\vec{X}$ back to $\vec{x}$ would be sufficient for yielding the effect ϕ. Finally, the minimality condition AC3 is used to rule out causally irrelevant conjuncts of an actual cause.[13]

Now, is an actual cause the 'genuine cause' we are looking for, when searching for a causal explanation? A variation of our forest fire example shows that this is not the case. Consider a situation in which we have $F = 1$, $A = 1$ and $L = 1$. But this time, the arsonist and the lightning strike are only jointly sufficient to produce a forest fire, which is expressed by the

[12] In case of a simple counterfactual dependence of ϕ on $\vec{X} = \vec{x}$, $\vec{W}$ is empty.

[13] Notice that Halpern and Pearl's definition (HP-definition) of actual causation formally allows for ϕ to be any Boolean combination of primitive events. All their examples, though, are restricted to cases, where the effect in question is a primtive event. Through most of this dissertation, I will make the same presupposition and assume that the effect or the explanandum in question is a primitive event. Whether Boolean combinations of primitive events are meaningful effects or explananda, is a question that I will leave largely untouched.

structural equation $F := A \wedge L$. It is easy to check that $A = 1$ and $L = 1$ are both actual causes of $F = 1$, since $F = 1$ counterfactually depends on $A = 1$ as well as on $L = 1$. Now imagine a detective, who wants to understand why the forest burns. She already knows the causal structure of the situation but not the real values of the variables except that $F = 1$. $L = 1$, despite being an actual cause of the explanandum, is no satisfying explanation for the detective, because she knows that a lightning strike alone is not enough to produce a forest fire. Since Halpern and Pearl aim to explicate a concept of complete or full explanation, they are looking for a cause that is actually sufficient for its effect. This is where the concept of sufficient causation comes into play. Here is how Halpern (2016, pp. 53–4) defines a sufficient cause:[14]

Sufficient Cause (Halpern, 2016). *$\vec{X} = \vec{x}$ is a 'sufficient cause' of ϕ in the causal setting $(\mathcal{M}, \vec{u})$ if and only if the following conditions hold:*

SC1 $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{x}$ *and* $(\mathcal{M}, \vec{u}) \models \phi$.

SC2 Some conjunct of $\vec{X} = \vec{x}$ is part of an actual cause of ϕ in $(\mathcal{M}, \vec{u})$. More precisely, there is a conjunct $X = x$ of $\vec{X} = \vec{x}$ and another (possibly empty) conjunction $\vec{Y} = \vec{y}$ such that $X = x \wedge \vec{Y} = \vec{y}$ is an actual cause of ϕ in $(\mathcal{M}, \vec{u})$; that is AC1, AC2 and AC3 hold for $X = x \wedge \vec{Y} = \vec{y}$.

SC3 $(\mathcal{M}, \vec{u}') \models [\vec{X} \leftarrow \vec{x}]\phi$ *for all contexts* $\vec{u}'$.

SC4 $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies conditions SC1, SC2 and SC3, where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}'$.

Again, SC1 says that for $\vec{X} = \vec{x}$ to be a sufficient cause of ϕ in $(\mathcal{M}, \vec{u})$, both $\vec{X} = \vec{x}$ and ϕ have to be true in $(\mathcal{M}, \vec{u})$. SC2 demands that some conjunct of the sufficient cause of ϕ is part of an actual cause of ϕ in $(\mathcal{M}, \vec{u})$.[15] Notice that it would be too strict to demand that a sufficient cause of ϕ in $(\mathcal{M}, \vec{u})$ must be an actual cause of ϕ in $(\mathcal{M}, \vec{u})$. Take the conjunctive forest fire scenario with $F := A \wedge L$ as an example. It is intuitively clear that $L = 1 \wedge A = 1$ is a sufficient cause of $F = 1$. But because of AC3, $L = 1 \wedge A = 1$ is not an actual cause of $F = 1$, since each conjunct of $L = 1 \wedge A = 1$ is an actual cause of $F = 1$.

SC3 is the condition that represents the essential intuition regarding sufficient causation. A sufficient cause must be sufficient for its effect ϕ in all contexts for the causal model $\mathcal{M}$. So, no matter what conditions $\vec{u}'$ hold, if the sufficient cause is true in $(\mathcal{M}, \vec{u}')$, then its effect ϕ is also true in $(\mathcal{M}, \vec{u}')$.[16] This is why in the conjunctive forest fire scenario neither $L = 1$ nor $A = 1$ is a sufficient cause of $F = 1$ on its own. SC4 is again a minimality condition, to get rid of causally irrelevant conjuncts.

[14] The definition of a sufficient cause as given in (Halpern and Pearl, 2005b) is quite unintuitive as a concept on its own. It makes sense only when embedded in the authors' definition of explanation. This is why I adopt Halpern's (2016) definition of sufficient causation and his reformulation of the HP-definition of explanation.

[15] I say that $\vec{X} = \vec{x}$ is a part of $\vec{Y} = \vec{y}$ if and only if all conjuncts of $\vec{X} = \vec{x}$ are conjuncts of $\vec{Y} = \vec{y}$.

[16] Notice, that the structural equations of a causal model often represent ceteris paribus (cp) laws, which only hold under the supposition of certain cp-conditions. In a setting of a causal model that presupposes the fulfilment of certain cp-conditions without representing them as variables an event might turn out as a sufficient cause of an effect even if it is no sufficient cause relative to a causal model that incorporates those cp-conditions as variables without presupposing their fulfilment.

1.3.2 Explanations and Epistemic States

We can now outline the basic idea of the HP-definition of explanation by the following slogan: an explanation, if found to be true, would constitute a sufficient cause of the explanandum. Notice that Halpern and Pearl do not simply equate explanations with sufficient causes. Most notably, they consider sufficient causes to be true formulas in a given causal setting, while they allow explanations to turn out as false. This is because Halpern and Pearl are not trying to explicate the concept of a correct explanation but the concept of a potential explanation (Halpern, 2016, p. 190). The idea of a potential explanation may seem strange, though. The explaining event either happened or it did not and if it did not happen, it cannot causally explain anything. So what exactly is the potentiality in potential explanations? To see this, we have to acknowledge that Halpern and Pearl are explicating an agent-dependent, epistemically relative conception of explanation, which is based on the understanding that the relation of being an explanation of something only exists for a certain agent α in a certain epistemic state $\mathcal{K}$.[17]. With this in mind, we can reformulate the above mentioned slogan a bit more precisely: the information $\vec{X} = \vec{x}$ is a potential explanation of ϕ for agent α if α is uncertain whether $\vec{X} = \vec{x}$ is true or false, but α believes that if $\vec{X} = \vec{x}$ is true, it would be a sufficient cause of ϕ. So, the potentiality of a potential explanation $\vec{X} = \vec{x}$ of ϕ for α is due to the epistemic uncertainty of α about whether $\vec{X} = \vec{x}$ is true.

Assuming an epistemic relativity of explanations is not new. Bas van Fraassen famously argued that "the term 'explains' is radically context-dependent" (van Fraassen, 1980, p. 91), in the sense that whether something counts as an explanation depends on the beliefs and concerns of the person who seeks the explanation. Halpern and Pearl follow the same sentiment. According to their account, whether $\vec{X} = \vec{x}$ is an explanation of ϕ for agent α depends on α's beliefs about the causal relation between $\vec{X} = \vec{x}$ and ϕ. But Halpern and Pearl (2005b, p. 897) add a further requirement that adds to the context dependency of explanations. They demand that an explanation for agent α should provide α only with information that goes beyond α's current beliefs. This demand is supported by the intuition that an explanation should enhance one's understanding, which presupposes a change in one's epistemic state. To summarize Halpern and Pearl's demands on explanations, we can formulate the following guideline:

> **Explanation Production Guideline (EPG).** To create an explanation of ϕ for agent α with epistemic state $\mathcal{K}$ take any $\vec{X} = \vec{x}$ about which α believes that, if it is true, it would be a sufficient cause of ϕ. Then remove any conjunct of $\vec{X} = \vec{x}$ that α already believes.

We have already equipped ourselves with a formal definition of sufficient causation. But for the explication of an epistemically relative concept of explanation, we still have to clarify how to represent epistemic states.

As pointed out above, the relations of actual and sufficient causation are always relative to a given causal setting which typically represents a very small snippet of the causal structure of the world. So, to assess whether some event $\vec{X} = \vec{x}$ is an actual or sufficient cause of another event ϕ one does not have to adopt a global perspective that takes the complete causal structure of the world into account. Instead, one considers an adequately localized representation $(\mathcal{M}, \vec{u})$

[17] See, for example, (Halpern, 2016, p. 188).

of the real-life causal scenario $\mathfrak{S}$ that encompasses the events $\vec{X} = \vec{x}$ and ϕ and determines whether $\vec{X} = \vec{x}$ fulfills the conditions of actual or sufficient causation relative to ϕ in $(\mathcal{M}, \vec{u})$. Halpern and Pearl adopt a correspondingly localized approach to causal explanations. To asses whether $\vec{X} = \vec{x}$ is an explanation of ϕ for an agent α one does not have to take into account the complete epistemic state of α. Instead, one only needs to consider the beliefs of α that concern the locally limited causal scenario $\mathfrak{S}$ that encompasses $\vec{X} = \vec{x}$ and ϕ. We can therefore model the relevant fragment of α's epistemic state as a set of causal settings that α considers to be potentially adequate representations of $\mathfrak{S}$. For reasons of simplicity, I will presuppose that there is never any uncertainty for α about which variables best represent a causal scenario $\mathfrak{S}$ and which structural equations hold in $\mathfrak{S}$. This means that all causal settings that α considers to be potentially adequate representations of $\mathfrak{S}$ have the same causal model $\mathcal{M}$ and differ only in their contexts $\vec{u}$. So, instead of representing α's epistemic state as a set of causal settings that describe $\mathfrak{S}$, we can simply represent it as a set of contexts for the causal model $\mathcal{M}$.[18] Attributions of degrees of belief can be added when needed.

1.3.3 The HP-Definition of Explanation

We are now equipped with all the relevant tools to present the HP-definition as given by Halpern (2016, pp. 188–9):

HP-Definition of Explanation (Halpern, 2016). *$\vec{X} = \vec{x}$ is an 'explanation' of ϕ relative to a set $\mathcal{K}$ of contexts in a causal model $\mathcal{M}$ if and only if the following conditions hold:*

EX1 $\vec{X} = \vec{x}$ is a sufficient cause of ϕ in all contexts in $\mathcal{K}$ that satisfy $(\vec{X} = \vec{x}) \wedge \phi$. More precisely:

(a) If $\vec{u} \in \mathcal{K}$ and $(\mathcal{M}, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \phi$, then there is a conjunct $X = x$ of $\vec{X} = \vec{x}$ and a (possibly empty) conjunction $\vec{Y} = \vec{y}$ such that $X = x \wedge \vec{Y} = \vec{y}$ is an actual cause of ϕ in $(\mathcal{M}, \vec{u})$.

(b) $(\mathcal{M}, \vec{u}') \models [\vec{X} \leftarrow \vec{x}]\phi$ for all contexts $\vec{u}' \in \mathcal{K}$.

EX2 $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies (EX1), where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}'$.

EX3 $\mathcal{K}_{(\vec{X}=\vec{x})\wedge\phi} \neq \varnothing$ (with $\mathcal{K}_{\psi} = \{\vec{u} \in \mathcal{K} : (\mathcal{M}, \vec{u}) \models \psi\}$).

The explanation is 'nontrivial' if it satisfies in addition:

EX4 $(\mathcal{M}, \vec{u}') \models \neg(\vec{X} = \vec{x})$ for some $\vec{u}' \in \mathcal{K}_{\phi}$.

[18] Halpern and Pearl (2005b) also consider the more general case, in which structural equations are considered to be unknown. They present a generalized definition of explanation for this case (Halpern and Pearl, 2005b, p. 907). Although I confine myself to scenarios, in which any uncertainty about structural equations is precluded, the results of this chapter can easily be expanded to the more general definition of explanation as well. However, an example that includes uncertainties about structural equations is typically significantly more complex than its counterpart that precludes such uncertainties.

As already pointed out, Halpern and Pearl aim for a notion of potential explanation, which means that agent α is uncertain about the truth of the explanation candidate. This is why EX4 demands that we only call an explanation $\vec{X} = \vec{x}$ nontrivial if the agent does not already believe that $\vec{X} = \vec{x}$ is true. But while it is not too bad if an agent already believes that the explanation is true (a trivial explanation is at least still an explanation), it is impermissible if the agent already believes that the explanation candidate is false. If α believes that $\vec{X} = \vec{x}$ did not happen, it cannot be a cause of anything according to α, which means that it cannot causally explain anything for α. Similarly, it makes no sense to explain an event that, according to α's beliefs, did not happen. This is why condition EX3 demands that the explanation $\vec{X} = \vec{x}$ as well as the explanandum ϕ is considered possible by α. EX2 is a minimality condition to get rid of irrelevant parts of an explanation. As we will see later, EX2 also plays a part in implementing the second instruction of EPG, namely to weed out those parts of an explanation that are already believed by α. The main purpose of EX1 is to implement the first instruction of EPG. But although Halpern's initial summary of EX1 suggests otherwise, the condition does not simply demand of an explanation $\vec{X} = \vec{x}$ to be a sufficient cause of ϕ in all causal settings $(\mathcal{M}, \vec{u})$ with $\vec{u} \in \mathcal{K}$ and $(\mathcal{M}, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \phi$. This becomes obvious when we compare the more precise formulation under EX1(a) and EX1(b) with the definition of a sufficient cause. While Condition EX1(a) corresponds perfectly with condition SC2, EX2(b) is actually a weakening of SC3, because the explanation $\vec{X} = \vec{x}$ is not required to be sufficient for ϕ in every logically possible causal setting, but only in the causal settings $(\mathcal{M}, \vec{u}')$ with $\vec{u}' \in \mathcal{K}$. I will call this weakened form of a sufficient cause a *sufficient cause modulo* $\mathcal{K}$:[19]

Sufficient Cause Modulo $\mathcal{K}$. *$\vec{X} = \vec{x}$ is a 'sufficient cause modulo $\mathcal{K}$' of ϕ in the causal setting $(\mathcal{M}, \vec{u})$ relative to a set $\mathcal{K}$ of contexts $\mathcal{M}$ if and only if the following conditions hold:*

SCM1 $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{x}$ *and* $(\mathcal{M}, \vec{u}) \models \phi$.

SCM2 Some conjunct of $\vec{X} = \vec{x}$ is part of an actual cause of ϕ in $(\mathcal{M}, \vec{u})$. More precisely, there is a conjunct $X = x$ of $\vec{X} = \vec{x}$ and another (possibly empty) conjunction $\vec{Y} = \vec{y}$ such that $X = x \wedge \vec{Y} = \vec{y}$ is an actual cause of ϕ in $(\mathcal{M}, \vec{u})$; that is AC1, AC2 and AC3 hold for $X = x \wedge \vec{Y} = \vec{y}$.

SCM3 $(\mathcal{M}, \vec{u}') \models [\vec{X} \leftarrow \vec{x}]\phi$ *for all contexts* $\vec{u}' \in \mathcal{K}$.

SCM4 $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies conditions SCM1, SCM2 and SCM3, where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}'$.

Having defined this concept of causation explicitly, we can also present the HP-definition of explanation in the following simplified form:

[19] I want to thank an anonymous referee for *The British Journal for the Philosophy of Science* for suggesting this name and for recommending to state the definition, which underlies Halpern's (2016) definition of explanation, separately.

HP-Definition of Explanation. *$\vec{X} = \vec{x}$ is an 'explanation' of ϕ relative to a set $\mathcal{K}$ of contexts in a causal model $\mathcal{M}$ if the following conditions hold:*

EX1' $\vec{X} = \vec{x}$ is a sufficient cause modulo $\mathcal{K}$ of ϕ in all contexts in $\mathcal{K}$ that satisfy $(\vec{X} = \vec{x}) \wedge \phi$.

EX3 $\mathcal{K}_{(\vec{X}=\vec{x})\wedge\phi} \neq \varnothing$ (with $\mathcal{K}_\psi = \{\vec{u} \in \mathcal{K} : (\mathcal{M}, \vec{u}) \models \psi\}$).

The explanation is 'nontrivial' if it satisfies in addition:

EX4 $(\mathcal{M}, \vec{u}') \models \neg(\vec{X} = \vec{x})$ for some $\vec{u}' \in \mathcal{K}_\phi$.

Halpern (2016) uses sufficient causation modulo $\mathcal{K}$ instead of sufficient causation simpliciter to ensure the compliance with the second instruction of EPG. For illustration, consider an example that is based on the conjunctive forest fire scenario, where the causal model $\mathcal{M}^F$ entails the following structural equations:

- $A := U_1$
- $L := U_2$
- $F := A \wedge L$

The causal diagram looks just like in Figure 1. Clearly, $A = 1 \wedge L = 1$ is a sufficient cause of $F = 1$ in every causal setting $(\mathcal{M}^F, \vec{u})$ in which $A = 1 \wedge L = 1$ holds. But there is no causal setting where $A = 1$ or $L = 1$ is a sufficient cause of $F = 1$, because neither $A = 1$ nor $L = 1$ can fulfill condition SC3. We now consider an epistemic state $\mathcal{K}$, in which the agent is uncertain whether $A = 1$ or $L = 1$ holds:

- $\mathcal{K} = \{\vec{u_0} = (1, 1), \vec{u_1} = (0, 1), \vec{u_2} = (1, 0), \vec{u_3} = (0, 0)\}$, with $\vec{u} = (u_1, u_2)$.[20]

It is easy to check that, according to the HP-definition, $A = 1 \wedge L = 1$ is an explanation of $F = 1$ in $\mathcal{M}$ relative to $\mathcal{K}$ while $A = 1$ and $L = 1$ alone are not. But the situation changes if we consider the epistemic state $\mathcal{K}'$ in which the agent believes $L = 1$:

- $\mathcal{K}' = \{\vec{u_0} = (1, 1), \vec{u_1} = (0, 1)\}$, with $\vec{u} = (u_1, u_2)$.

$A = 1$ is now an explanation of $F = 1$ in $\mathcal{M}$ relative to $\mathcal{K}'$ according to the HP-definition, while EX2 makes sure that $A = 1 \wedge L = 1$ is not. So, $A = 1$ can be an explanation of $F = 1$ without being a sufficient cause of $F = 1$ in any causal setting. This is possible because the HP-definition employs the concept of sufficient causation modulo $\mathcal{K}$ instead of the concept of sufficient causation simpliciter: To be an explanation of $F = 1$ in the given example, $A = 1$ does not need to be sufficient for $F = 1$ in all logically possible contexts for $\mathcal{M}$. It only needs to be sufficient for $F = 1$ in all causal settings $(\mathcal{M}, \vec{u})$ with $\vec{u} \in \mathcal{K}'$. This is how the HP-definition satisfies the second demand of EPG and removes $L = 1$ from the explanation, in case that the recipient of the explanation already believes that $L = 1$ holds.

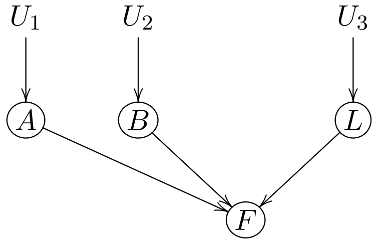
[20] Here and in the following representations of epistemic states, the elements $\vec{u_i}$ of $\mathcal{K}$ represent contexts which are sequences of values of the exogenous variables of the causal model under consideration. The first entry in a sequence $\vec{u_i}$ represents the value u_1 of the exogenous variable U_1, the second entry represents the value u_2 of U_2, and so on.

1.4 Counterexamples to the HP-Definition of Explanation

In the conjunctive forest fire scenario, as well as in several other examples,[21] the HP-definition perfectly abides by EPG and delivers intuitively sound results. I will now provide examples to show that this is not always the case.

1.4.1 The Overdetermined Forest Fire

The first example is a slight variation of the forest fire scenario. A lightning strike $L = 1$ is now sufficient for a forest fire. But the arsonist can also cause a forest fire without a lightning strike if he not only drops a lit match ($A = 1$) but additionally spills some benzine ($B = 1$). This gives us the causal model $\mathcal{M}^{OF}$ as presented in figure 1.2.



- $A := U_1$
- $B := U_2$
- $L := U_3$
- $F := (A \wedge B) \vee L$

Figure 1.2: $\mathcal{M}^{OF}$ - the overdetermined forest fire scenario.

We consider the following epistemic state $\mathcal{K}^{OF}$, which includes the belief that the lightning strike took place and the belief that no arsonist drops a match without spilling benzine:

- $\mathcal{K}^{OF} = \{\vec{u_0} = (1,1,1), \vec{u_1} = (0,1,1), \vec{u_2} = (0,0,1)\}$, with $\vec{u} = (u_1, u_2, u_3)$.

According to the HP-definition, $A = 1$ is a nontrivial explanation of $F = 1$ relative to $\mathcal{K}^{OF}$ in $\mathcal{M}^{OF}$.

Proof. EX1(a) is satisfied if for all $\vec{u} \in \mathcal{K}^{OF}$ with $(\mathcal{M}^{OF}, \vec{u}) \models A = 1 \wedge F = 1$ there is a $\vec{Y} = \vec{y}$ such that $A = 1 \wedge \vec{Y} = \vec{y}$ is an actual cause of $F = 1$ in $(\mathcal{M}^{OF}, \vec{u})$. $\vec{u_0}$ is the only context in $\mathcal{K}^{OF}$ with $(\mathcal{M}^{OF}, \vec{u_0}) \models A = 1 \wedge F = 1$. Take $\vec{Y} = \varnothing$, then AC1 is satisfied, since $(\mathcal{M}^{OF}, \vec{u_0}) \models A = 1 \wedge F = 1$. Take $\vec{W} = \{L\}$ and $\vec{w} = 0$, then AC2(a) is satisfied, because $(\mathcal{M}^{OF}, \vec{u_0}) \models [A \leftarrow 0, L \leftarrow 0]F = 0$, and AC2(b) is satisfied, because $(\mathcal{M}^{OF}, \vec{u_0}) \models [A \leftarrow 1, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]F = 1$ for all $\vec{W}' \subseteq \vec{W}$ and for all $\vec{Z}' \subseteq \vec{Z}$ (with $A = 1$ and $B = 1$, we have $F = 1$, no matter the value of L). $A = 1$ clearly satisfies the minimality condition AC3. So, $A = 1$ is an actual cause of $F = 1$ in $(\mathcal{M}^{OF}, \vec{u_0})$, which means that EX1(a) is satisfied. EX1(b) is satisfied, because $(\mathcal{M}^{OF}, \vec{u}') \models [A \leftarrow 1]F = 1$ for all contexts $\vec{u}' \in \mathcal{K}^{OF}$. $A = 1$ also clearly satisfies the minimality condition EX2. EX3 is satisfied, since $\mathcal{K}^{OF}_{A=1 \wedge F=1} = \{\vec{u_0}\} \neq \varnothing$. EX4 is satisfied, because of $\vec{u_1}$ and $\vec{u_2}$. □

[21] See, for example, (Halpern, 2016, pp. 189–92).

Notice that $A = 1$ is considered to be an explanation of $F = 1$ by the HP-definition, although $A = 1$ is no sufficient cause of $F = 1$ in any causal setting of $\mathcal{M}^{OF}$. On the other hand, $A = 1 \wedge B = 1$ is a sufficient cause of $F = 1$ in any causal setting of $\mathcal{M}^{OF}$ in which it is true. But, because of EX2, the HP-definition does not recognize it as an explanation of $F = 1$. According to EPG, this would be fine, if $B = 1$ would already be believed in $\mathcal{K}^{OF}$. In that case, $A = 1 \wedge B = 1$ would contain information that is redundant for an agent α in epistemic state $\mathcal{K}^{OF}$ and $A = 1$ would suffice for a full explanation. But $B = 1$ is not believed in $\mathcal{K}^{OF}$, which means that the HP-definition deviates from EPG in this example.

But does this speak against the HP-definition? One could argue that although α does not believe $B = 1$ in her current epistemic state, she ideally could and rationally should believe $B = 1$ right after learning $A = 1$ because $B = 1$ logically follows from her background beliefs $\mathcal{K}^{OF}$ and $A = 1$. This is why, one could claim, the information $B = 1$ is redundant for α. If we accept this defence of the HP-definition, we automatically pledge allegiance to a stricter version of EPG:

> **EPG'.** To create an explanation of ϕ for agent α with epistemic state $\mathcal{K}$ take any $\vec{X} = \vec{x}$ about which α believes that, if it is true, it would be a sufficient cause of ϕ. Then remove any conjunct of $\vec{X} = \vec{x}$ that is deducible from $\mathcal{K}$ and the remaining conjuncts of $\vec{X} = \vec{x}$.

EPG' certainly is a fitting guideline for constructing explanations for logical omniscient agents. But logical omniscience is a much too high standard for any human being.[22] In the context of human endeavors it is therefore highly unrealistic and impractical to presuppose logical omniscience and to dispose of any information that is deemed redundant only because it is in principle deducible. Instead of producing explanations that spare agents from already believed information, such an approach tends to confront human agents with riddles by only presenting the smallest possible amount of clues by which an agent can in principle deduce the full explanation herself.

But note that the HP-definition is by no means faithful to EPG' either. Consider an epistemic state, which does not include a belief in $L = 1$ but which still includes the belief that no arsonist drops a match without spilling benzine:

- $\mathcal{K}'^{OF} = \{\vec{u_0} = (1, 1, 1), \vec{u_1} = (0, 1, 1), \vec{u_2} = (0, 0, 1), \vec{u_3} = (1, 1, 0), \vec{u_4} = (0, 0, 0)\}$, with $\vec{u} = (u_1, u_2, u_3)$.

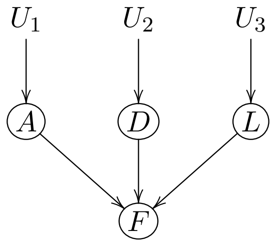
Now, the HP-definition does not recognize $A = 1$ as an explanation of $F = 1$ relative to $\mathcal{K}'^{OF}$ in $\mathcal{M}^{OF}$, although it is still the case that $B = 1$ is deducible from $A = 1$ and $\mathcal{K}'^{OF}$.[23] So, even for someone who might prefer EPG' over EPG, the HP-definition is no good choice. It alternates between following EPG and following EPG' and the reason it does so, namely whether the agent believes in the presence of another sufficient cause of the explanandum, seems rather arbitrary.

[22] See, for example, (Hintikka, 1962, pp. 30–31).

[23] $A = 1$ does not fulfill EX1(b) because of $\vec{u_4}$.

1.4.2 The Dry Forest

For the next example we change the forest fire scenario once again. Consider a forest in a rainy country. A necessary condition for a fire is that the forest is sufficiently dry. Since we cannot simply presuppose that condition anymore, we add the variable D to the model, which represents whether the forest is dry. Now, an arsonist and a dry forest are jointly sufficient for a forest fire and, similarly, a lightning strike and a dry forest are jointly sufficient for a forest fire. This gives us the causal model $\mathcal{M}^D$ as presented in figure 1.3.



- $A := U_1$
- $D := U_2$
- $L := U_3$
- $F := (A \wedge D) \vee (D \wedge L)$

Figure 1.3: $\mathcal{M}^D$ - the dry forest scenario.

We consider an agent α who already believes that the forest is burning. She further believes that the fire was either caused by the arsonist combined with the dry forest or by a lightning strike combined with the dry forest. So we have:

- $\mathcal{K}^D = \{\vec{u_0} = (1, 1, 0), \vec{u_1} = (0, 1, 1)\}$, with $\vec{u} = (u_1, u_2, u_3)$.

According to the HP-definition, $D = 1$ is a trivial explanation of $F = 1$ relative to $\mathcal{K}^D$ in $\mathcal{M}^D$.[24] Therefore, the HP-definition does not recognize $A = 1 \wedge D = 1$ nor $D = 1 \wedge L = 1$ as an explanation of $F = 1$ relative to $\mathcal{K}^D$ in $\mathcal{M}^D$ because of minimality. Here again, $D = 1$ is no sufficient cause of $F = 1$ in any causal setting of $\mathcal{M}^D$, while $D = 1 \wedge A = 1$ and $D = 1 \wedge L = 1$ are sufficient causes of $F = 1$ in any causal setting of $\mathcal{M}^D$, in which they are true. Since neither $A = 1$ nor $L = 1$ is believed in $\mathcal{K}^D$, this example illustrates another deviation of the HP-definition from EPG. But in contrast to the previous example, neither $A = 1$ nor $L = 1$ is deducible from $D = 1$ and $\mathcal{K}^D$.

To defend the HP-definition, one could argue that acknowledging $D = 1$ as an explanation of $F = 1$ is fair enough, because the agent believes that there is a sufficient cause of $F = 1$ of which $D = 1$ is a part, even if she cannot use the information $D = 1$ to identify the entire sufficient cause. Lewis (1986c), for example, argues that any information about the causal history of an event counts as an explanation. But these are very low demands, especially when we consider that the HP-definition is meant to explicate a conception of full explanation.[25]

[24] It is easy to see that for all $\vec{u} \in \mathcal{K}^D$ with $(\mathcal{M}^D, \vec{u}) \models D = 1 \wedge F = 1$, $D = 1$ is an actual cause of $F = 1$ in $(\mathcal{M}^D, \vec{u})$. Furthermore, $D = 1$ is sufficient for $F = 1$ in every $(\mathcal{M}^D, \vec{u})$ with $\vec{u} \in \mathcal{K}^D$. Finally, $D = 1$ is obviously minimal.

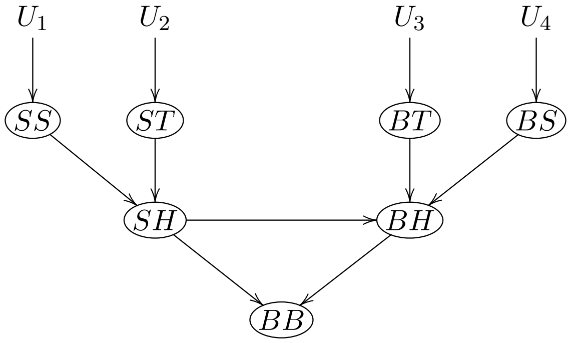
[25] As we will see later, Halpern and Pearl (2005b, p. 904) also define partial explanations, but they use the concept of a full explanation as defined in the HP-definition to do so.

But even Lewis admits that more information about the causal history of an event amounts to a better explanation of that event. So, it is a very odd result that, due to the minimality condition, the HP-definition does not acknowledge anything as an explanation that entails more information besides $D = 1$, particularly $D = 1 \wedge A = 1$ and $D = 1 \wedge L = 1$.

A more demanding and much more coherent approach to complete causal explanations would ask of any explanation of ϕ for agent α to be such that if α believes the content of the explanation to be true, α considers herself to be able to identify a unique sufficient cause of ϕ which is responsible for ϕ's production. This demand is not just much more in line with EPG, but also with several mechanistic accounts of causal explanation, according to which an explanation has to point out the mechanism that is responsible for the production of the explanandum.[26] According to this view, whenever an agent asks why $F = 1$ happened, while already believing that $F = 1$ is caused either by $D = 1 \wedge A = 1$ or by $D = 1 \wedge L = 1$, the answer 'because $D = 1$' does not satisfy the search for a complete causal explanation in any way.

1.4.3 Suzy, Billy, and the Bottle

As a final example, consider the causal model $\mathcal{M}^{SB}$ as presented in figure 1.4. It represents a variant of the bottle breaking scenario with Suzy and Billy that I briefly introduced above.



- $SS := U_1$
- $ST := U_2$
- $BT := U_3$
- $BS := U_4$

- $SH := SS \wedge ST$
- $BH := (BT \wedge BS) \wedge \neg SH$
- $BB := SH \vee BH$

Figure 1.4: $\mathcal{M}^{SB}$ - the bottle breaking scenario.

The epistemic state is:

- $\mathcal{K}^{SB} = \{\vec{u_0} = (1,1,1,1), \vec{u_1} = (0,1,1,1), \vec{u_2} = (1,1,0,1), \vec{u_3} = (0,0,1,1), \vec{u_4} = (1,0,1,1)\}$, with $\vec{u} = (u_1, u_2, u_3, u_4)$.

Both Billy and Suzy are considered to be perfect stone-throwers as long as they are sober, otherwise they always miss. So, if Suzy throws a stone at a bottle ($ST = 1$) and Suzy is sober

[26] See, for example, (Machamer et al., 2000).

($SS = 1$), she will definitely hit the bottle ($SH = 1$). The bottle breaks ($BB = 1$) if either Billy or Suzy hits the bottle ($SH = 1 \vee BH = 1$). But since Suzy throws a bit harder than Billy, her stone will hit and break the bottle before Billy's stone has a chance of hitting the bottle. So, if Billy throws ($BT = 1$) and he is sober ($BS = 1$), he will only hit the bottle ($BH = 1$), if Suzy did not hit it already ($SH = 0$).

According to the HP-definition, neither $BT = 1$ nor $ST = 1$ is an explanation of $BB = 1$ relative to $\mathcal{K}^{SB}$ in $\mathcal{M}^{SB}$, since there are contexts in $\mathcal{K}^{SB}$, in which $BT = 1 \wedge BB = 1$ is the case, while $BT = 1$ is no part of an actual cause of $BB = 1$ (this is the case in context $\vec{u_0}$), or in which $ST = 1 \wedge BB = 1$ is the case, while $ST = 1$ is no part of an actual cause of $BB = 1$ (this is the case in context $\vec{u_1}$). But, according to the HP-definition, $ST = 1 \wedge BT = 1$ is an explanation of $BB = 1$ relative to $\mathcal{K}^{SB}$ in $\mathcal{M}^{SB}$.

Proof. EX1(a) is satisfied if for all $\vec{u} \in \mathcal{K}^{SB}$ with $(\mathcal{M}^{SB}, \vec{u}) \models ST = 1 \wedge BT = 1 \wedge BB = 1$, namely $\vec{u_0}$ and $\vec{u_1}$, there is a conjunct of $ST = 1 \wedge BT = 1$ which is part of an actual cause of $BB = 1$ in $(\mathcal{M}^{SB}, \vec{u})$. In $(\mathcal{M}^{SB}, \vec{u}_0)$, $ST = 1$ is part of an actual cause of $BB = 1$ and in $(\mathcal{M}^{SB}, \vec{u}_1)$, $BT = 1$ is part of an actual cause of $BB = 1$. EX1(b) is satisfied because $(\mathcal{M}^{SB}, \vec{u}') \models [ST \leftarrow 1, BT \leftarrow 1]BB = 1$ holds for all $\vec{u}' \in \mathcal{K}^{SB}$. EX2 is satisfied because no smaller part of $ST = 1 \wedge BT = 1$ fulfills EX1 relative to $\mathcal{K}^{SB}$. EX3 is satisfied, since $\mathcal{K}^{SB}_{ST=1 \wedge BT=1 \wedge BB=1} = \{\vec{u}_0, \vec{u}_1\} \neq \varnothing$. EX4 is satisfied because of $\vec{u}_2$, $\vec{u}_3$, and $\vec{u}_4$. □

If α with epistemic state $\mathcal{K}^{SB}$ and knowledge about the causal relations in $\mathcal{M}^{SB}$ asks why the bottle broke, it may first seem like a good explanation to say that Billy and Suzy both threw a stone at it. The answer provides α with new information and it makes the explanandum much less surprising. The explanandum $BB = 1$ is even deducible from $\mathcal{K}^{SB}$ and $ST = 1 \wedge BT = 1$, which makes it a good specimen of Hempel and Oppenheim's (1948) deductive-nomological (DN) conception of explanation. But it also embodies some of the substantial problems of the DN-account, namely its neglect of actual causes of the explanandum in preemption scenarios and its inability to preclude causally irrelevant factors. Even if α believes that $ST = 1 \wedge BT = 1$ is true and she accepts it as an explanation of $BB = 1$, she is, just like in the previous example, still not able to tell which event actually caused the explanandum and is therefore responsible for the bottle breaking. So, $ST = 1 \wedge BT = 1$ does not entail enough information to be a satisfying causal explanation according to EPG. But at the same time, it entails too much information, since for every $\vec{u} \in \mathcal{K}^{SB}$ with $(\mathcal{M}^{SB}, \vec{u}) \models ST = 1 \wedge BT = 1 \wedge BB = 1$ the explanation $ST = 1 \wedge BT = 1$ contains a part that is causally irrelevant for the explanandum. In $\vec{u_0}$, $BT = 1$ is causally irrelevant for the explanandum because $SS = 1 \wedge ST = 1$ is the sufficient cause of $BB = 1$, which preempts any causal influence from $BT = 1$ on the value of BB. And in $\vec{u_1}$, $ST = 1$ is causally irrelevant for the explanandum because $SS = 0$ holds, which prevents any causal influence from $ST = 1$ on the value of BB. So α is aware that, no matter which context is actually true, at least one conjunct of the explanation $ST = 1 \wedge BT = 1$ is causally irrelevant to the explanandum.

The ability to preclude explanatory irrelevant information from explanations is commonly considered as a crucial requirement on any account of explanation. To say it in Salmon's words: 'irrelevancies [are] harmless to arguments but fatal to explanations' (Salmon, 1989, p. 102). For a conception of causal explanation, causal irrelevance typically amounts to explanatory irrelevance. It is therefore remarkable that there are scenarios, in which the HP-definition allows explanations that not only fail to identify what, if true, would be a sufficient cause of the explanandum, but also entail causally irrelevant information. It also amounts to another violation of EPG, which demands to name something that, if true, would be a sufficient cause, which by definition does not entail causally irrelevant parts. The second instruction of EPG only allows to remove certain information, but not to add causally irrelevant information.

The HP-definition yields another result in the given example that is at least dubious. While it adequately recognizes $ST = 1 \wedge SS = 1$ as a potential explanation of $BB = 1$, it does not consider $BS = 1 \wedge BT = 1$ to be a potential explanation of $BB = 1$. $BS = 1 \wedge BT = 1$ violates condition EX1, because in $\vec{u}_0$ no part of $BS = 1 \wedge BT = 1$ is an actual cause of $BB = 1$. This is the context in which Suzy's throw preempts any causal influence from $BS = 1 \wedge BT = 1$ to $BB = 1$. But I consider it to be rather strange to say that $BS = 1 \wedge BT = 1$ is no potential explanation of $BB = 1$, even though we know that it might very well be the case that $BS = 1 \wedge BT = 1$ causally produces $BB = 1$. Yes, the causal influence from $BS = 1 \wedge BT = 1$ to $BB = 1$ might be preempted by another cause. But remember that we are only considering whether $BS = 1 \wedge BT = 1$ is a *potential* explanation of $BB = 1$ and not whether it is an actual or a correct explanation of it. The fact that it might causally produce $BB = 1$ should be enough to qualify as a potential explanation, even though it is not enough for an actual or a correct explanation.

1.5 Amending the HP-Definition

In the previous section, I have shown that the HP-definition does not live up to its own demands which are summarized informally in EPG. As a consequence, instead of just sparing agents from already believed information, the HP-definition occasionally, but not always, excludes causally relevant information from an explanation only because an agent could in principle deduce the information by herself. Additionally, the HP-definition occasionally admits explanations for an agent α, that, even if they are true and accepted by α, leave α completely ignorant about which causal process is actually responsible for the production of the explanandum. And finally, the HP-definition occasionally admits explanations with causally irrelevant information. In this section, I will argue that it is possible to eliminate these flaws of the HP-definition while preserving all the benefits that come with it. To achieve this, I will start with re-considering on what kind of causal concept we should built our definition of explanation.

1.5.1 Strong Actual Causes

We have seen that the crucial condition AC2 of actual causation consists of two parts. The first, condition (a), expresses the idea that in a certain contingency the cause is necessary for the effect: without the cause, the effect would not have happened. The second condition, condition (b), expresses the idea that in that very same contingency (as well as in some other contingencies) the cause must also be sufficient for the effect, in the sense that realizing the cause by an intervention would yield the effect. We have already seen that the second condition, the sufficiency condition of actual causation, is too weak to make an actual cause a reasonable condidate for a complete causal explanation. Some actual causes do intuitively not qualify to count as complete potential explanations. This is why Halpern and Pearl have put the emphasis on the concept of sufficient causation for explicating complete potential explanations. While it retains the condition of necessity AC2(a), it crucially strengthens the condition of sufficiency by adding condition SC3: the cause must must be sufficient for the effect in every logically possible context for the causal model in question. But we have seen that this sufficiency condition turned out to be too strong for explicating a concept of potential explanation that is in line with EPG. Some events seem to be complete potential explanations, even though they do not satisfy the sufficiency-requirement of sufficient causation.

As pointed out above, Halpern therefore based his definition of potential explanation on a different concept of causation, namely sufficient causation modulo $\mathcal{K}$. This concept rests on a sufficiency condition that lies somewhere in between the weak sufficiency condition for actual causes (AC2(b)) and the strong sufficiency condition for sufficient causes (SC3): SCM3 demands that realizing the cause by an intervention must yield the effect in all contexts for the given causal model $\mathcal{M}$ that are in α's epistemic state $\mathcal{K}$. I have argued in the previous section that building a concept of explanation on the concept of sufficient causation modulo $\mathcal{K}$ leads to serious problems. But on which concept of causation should we build our explications of explanation, if the concept of actual causation is too weak, the concept of sufficient causation is too strong and the concept of sufficient causation modulo $\mathcal{K}$ has all the downsides that we have just uncovered.

Luckily, there is still an alternative, albeit one that has not got much attention in the literature on causation. This concept of causation, which I will call *strong actual causation*, has been considered and discussed, more or less briefly, by Pearl (2000, p. 317) (under the name of *sustenance*), Halpern and Pearl (2005a, p. 855) (under the name of *strong causation*), and Weslake (2011) (under the name of *weakly sufficient actual causation*).[27] More recently, Beckers (2021) brings up the concept again, but only to mention that it has been widely neglected so far. Despite the fact that the concept of strong actual causation has not played any significant role in the literature on causation so far, I will show in the course of this dissertation that it is the most promising candidate for explicating concepts of causal explanation. The reason is basically this: The concept of strong actual causation is based on a sufficiency condition that is more demanding than the one of actual causation and in crucial respects less demanding than the one of sufficient causation. At the same time, it does not lead us into the same trouble as the concept of sufficient causation modulo $\mathcal{K}$.

To define strong actual causation, we first need to to explicate the concept of an *active causal path.* A *causal path* in a causal model $\mathcal{M}$ is an n-tuple $(V_i)_{i=1,\dots,n}$ consisting of endogenous variables in $\mathcal{M}$ with V_i being a parent of V_{i+1} for every i with $1 \leq i < n$. We will say that a causal path $(V_i)_{i=1,\dots,n}$ is *active* in a causal setting $(\mathcal{M}, \vec{u})$ if and only if $(\mathcal{M}, \vec{u}) \models V_i = v_i$ for every i with $1 \leq i < n$ and $V_i = v_i$ is an actual cause of $V_{i+1} = v_{i+1}$ in $(\mathcal{M}, \vec{u})$ for every i with $1 \leq i < n$. With this at hand, we can now formulate our definition of strong actual causation:

[27] The definitions of sustenance, strong causation, and weakly sufficient actual causation all differ in some respects. But they are still all similar enough to count as explications of one and the same concept.

Strong Actual Causation. *$\vec{X} = \vec{x}$ is a 'strong actual cause' of ϕ in the causal setting $(\mathcal{M}, \vec{u})$ (in short: $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{x} \rightarrowtail \phi$) if and only if the following conditions hold:*

SAC1 $(\mathcal{M}, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \phi$.

SAC2 There is a partition $(\vec{Z}, \vec{W})$ of the endogenous variables $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w})$ such that if $(\mathcal{M}, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then:

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

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

SAC3 Let $\vec{V}$ be all variables in $\mathcal{V}$ that are not in any causal path from a variable in $\vec{X}$ to a variable in ϕ that is active in $(\mathcal{M}, \vec{u})$, then: $(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{V} \leftarrow \vec{v}]\phi$ for all settings $\vec{v}$ of $\vec{V}$.

SAC4 $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ (with $\vec{x}'$ being the restriction of $\vec{x}$ to the variables in $\vec{X}'$) satisfies conditions SAC1, SAC2 and SAC3.

Conditions SAC1 and SAC2 are the same as for actual causation. A strong actual cause only has to satisfy one additional condition, which is SAC3. The minimality condition is then applied after SAC3, such that it has to consider the fulfillment of SAC3, when straining out non-minimal causes. The idea that underlies SAC3 is this: The strong actual cause $\vec{X} = \vec{x}$ always yields the effect ϕ in the given causal setting, no matter what values all other variables in the model are set to, except for those variables that lie on an active causal path from $\vec{X}$ to ϕ. The condition expresses a certain form of sufficiency: $\vec{X} = \vec{x}$ is not dependent on any other factors or events to causally produce the effect ϕ in the given situation, except, of course, the causal intermediaries between $\vec{X} = \vec{x}$ and ϕ. But notice that this form of sufficiency only has to hold in the given causal setting $(\mathcal{M}, \vec{u})$. This is the crucial difference to the sufficiency condition SC3 in the definition of sufficient causation, which demands that $[\vec{X} \leftarrow \vec{x}]\phi$ must be true in every logically possible causal setting for $\mathcal{M}$.

One might wonder, whether this sufficiency condition is still too strong for the concept of potential explanation in the spirit of EPG. Consider again the conjunctive forest fire scenario from section 1.3.3, where we have a causal model $\mathcal{M}$ with the following structural equations:

- $A := U_1$
- $L := U_2$
- $F := A \wedge L$

It is clear that, according to EPG, $A = 1$ should be acknowledged as a potential explanation of $F = 1$ in $\mathcal{M}$ relative to the epistemic state $\mathcal{K}' = \{\vec{u_0} = (1, 1), \vec{u_1} = (0, 1)\}$, in which $L = 1$ is already fully believed. But $A = 1$ is no strong actual cause of $F = 1$ in either $\vec{u_0}$ or $\vec{u_1}$, since in both contexts L does not lie on any active causal path from A to F and setting L on 0 ensures that $A = 1$ is not sufficient for $F = 1$. So it seems that the concept of strong actual causation has, just like the concept of sufficient causation, a sufficiency condition that is still too strong for explicating concepts of explanation that abide by EPG.

This worry is appropriate but it does not disqualify the concept of strong actual causation for explicating concepts of explanation. We will later see that, if we use the concept of strong actual causation in the right way for explicating concepts of causal explanation, then the problem that we have just singled out will not arise. But before we can get to this, we still have to do some preparatory work. So far, I have treated EPG as the undisputed guideline for defining explanation. But there are some issues with EPG that we have to address before amending the HP-definition of explanation.

1.5.2 Adjusting EPG

What Kind of Causation Is Needed for Explanation?

My first adjustment of EPG should already be clear from the previous section. EPG, as formulated in section 1.3.2 builds the concept of complete potential explanation on sufficient causation. Instead, I aim to build explanations on the concept of strong actual causation.

What Kind of Uncertainty Renders an Explanation Potential?

I have already stressed that EPG is supposed to be a guidance for explicating the concept of a complete *potential* explanation and not for the concept of a correct or an actual explanation. As discussed in section 1.3.2, Halpern and Pearl understand the potentiality of a potential explanation as an agent's uncertainty about whether the explanation candidate $\vec{X} = \vec{x}$ is actually the case in the given situation. This is why EPG, as formulated so far, does not demand that α already believes that $\vec{X} = \vec{x}$ is a genuine cause of ϕ. Instead it only demands that α believes the following: if $\vec{X} = \vec{x}$ is indeed the case, then $\vec{X} = \vec{x}$ is a genuine cause of ϕ. This formulation leaves room for α's uncertainty about whether $\vec{X} = \vec{x}$ is really the case. But I consider this to be an unintuitively narrow interpretation of the potentiality of a potential explanation. It seems more adequate to say that what makes an event $\vec{X} = \vec{x}$ a *potential* explanation of ϕ for a given agent α instead of an actual or a correct explanation is the fact that α is still uncertain, not about whether $\vec{X} = \vec{x}$ is the case or not, but about whether $\vec{X} = \vec{x}$ really is a genuine cause of ϕ or not. The first uncertainty entails the second, but not vice versa. Consider the event $BT = 1 \wedge BS = 1$ in the bottle breaking scenario. If you are uncertain about whether $BT = 1 \wedge BS = 1$ is the case, you are also uncertain about whether $BT = 1 \wedge BS = 1$ is a genuine cause of $BB = 1$. But one may be uncertain about whether $BT = 1 \wedge BS = 1$ is a genuine cause of $BB = 1$, while fully believing that $BT = 1 \wedge BS = 1$ is the case, since $BT = 1 \wedge BS = 1$, even if it is the case, might or might not be a genuine cause of $BB = 1$.[28]

I do not see any reason, why we should restrict the concept of a potential explanation to those explanation candidates about which α is uncertain whether it is actually the case or not. Intuitively, explanation candidates about which α is uncertain whether they really are genuine causes of the explanandum fit the description of potential explanations just as well. This is why I propose the following reformulation of EPG:

[28] For example because it might be unclear whether another cause preempts $BT = 1 \wedge BS = 1$.

> **Explanation Production Guideline (EPG).** To create a (complete potential) explanation of ϕ for agent α with epistemic state $\mathcal{K}$ take any $\vec{X} = \vec{x}$ that, according to α's epistemic state, might actually be a strong actual cause of ϕ. Then remove any conjunct of $\vec{X} = \vec{x}$ that α already believes.

Is Already Known Information Really Toxic for Explanations?

Is it really justified to demand that no explanation for α may entail information that α already believes? Is already believed information really as toxic for an explanation as causally irrelevant information? When I ask a ranger why there is a forest fire and she replies: 'There was a lightning strike and the owls did not sleep well last week', I do not regard her answer as an explanation of the forest fire. Instead, the ranger just utterly mystified me. By giving a causal explanation the ranger not only asserts the truth of every conjunct of her answer but also the causal relevance of every conjunct to the explanandum. If this relevance relation does not hold for one of the conjuncts, then the answer fails as an explanation. Now, imagine that the ranger says: 'There has been a lightning strike and the forest is very dry', when I already believe that the forest is dry. It seems very strange to say that the ranger's answer is not a satisfying explanation only because I already believe a part of her answer. It rather highlights that the explanation fits quite well into my background beliefs. So, it seems that informatively redundant conjuncts, unlike causally irrelevant ones, do not really harm the value of an explanation. But that means that the second instruction of EPG seems to be way too strict.

So, should we replace EPG with a guideline for a formal explication of *explanation* that is less ruthless with informatively redundant conjuncts? I think that depends on what we aim to do with the formal explication of *explanation*. If we want to use it as a blueprint for a program that produces potential explanations for an agent α after receiving α's epistemic state $\mathcal{K}$ and a query 'Why ϕ?' as inputs, then it makes sense to be economical and to only produce explanations that are as short as possible, which is exactly what EPG commends. But if we want to use the formal explication of *explanation* like an auditing agency which processes requests of the form: 'Please determine whether $\vec{X} = \vec{x}$ is an explanation of ϕ for agent α with epistemic state $\mathcal{K}$', then it seems inappropriately harsh to reject explanations which entail information that α already believes. Rather than banning the inclusion of information that α already believes, the auditing agency should permit the omission of information that α already believes. This gives us the following guideline for the recognition of explanations:

> **Explanation Recognition Guideline (ERG).** A (complete potential) explanation of ϕ for agent α with epistemic state $\mathcal{K}$ is any $\vec{X} = \vec{x}$ that according to $\mathcal{K}$ might be a strong actual cause of ϕ. Removing conjuncts of $\vec{X} = \vec{x}$ that α already believes does not change that assessment.

Different goals impose different requirements on explanations. In my amendment of the HP-definition I will take this into account by differentiating several conceptions of causal explanation.

1.5.3 The Amended Definitions of Explanation

I will first present the amended definitions in full. Afterwards I will discuss how they deviate from the HP-definition:

Potential Explanation. *$\vec{X}=\vec{x}$ is a 'potential explanation' of ϕ relative to a set $\mathcal{K}$ of contexts in a causal model $\mathcal{M}$ if and only if the following conditions hold:*

E1 There is an event $\vec{S}=\vec{s}$ with $\vec{X}=\vec{x}$ being a part of $\vec{S}=\vec{s}$ and $(\mathcal{M},\vec{u})\models\vec{S}=\vec{s}\rightarrowtail\phi$ for some $\vec{u}$ in $\mathcal{K}$.

E2 For all contexts $\vec{u}\in\mathcal{K}:(\mathcal{M},\vec{u})\models\vec{S}'=\vec{s}'$ (where $\vec{S}'=\vec{S}\setminus\vec{X}$ and $\vec{s}'$ is the restriction of $\vec{s}$ to the variables in $\vec{S}'$).

E3 $(\mathcal{M},\vec{u}')\models\neg(\vec{S}=\vec{s}\rightarrowtail\phi)$ for some $\vec{u}'\in\mathcal{K}_\phi$.

Actual Explanation. *$\vec{X}=\vec{x}$ is an 'actual explanation' of ϕ relative to $\mathcal{K}$ in $\mathcal{M}$ if and only if it satisfies E1, E2, and:*

E4 $(\mathcal{M},\vec{u})\models\vec{S}=\vec{s}\rightarrowtail\phi$ for all $\vec{u}\in\mathcal{K}$.

Parsimonious Potential Explanation. *A potential explanation $\vec{X}=\vec{x}$ of ϕ relative to $\mathcal{K}$ in $\mathcal{M}$ is 'parsimonious' if and only if it satisfies E1, E2, E3, and:*

E5 $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}'=\vec{x}'$ satisfies E1 and E2.

Explanation. *We will simply say that $\vec{X}=\vec{x}$ is an 'explanation' of ϕ relative to $\mathcal{K}$ in $\mathcal{M}$ if and only if $\vec{X}=\vec{x}$ is either a potential or an actual explanation of ϕ relative to $\mathcal{K}$ in $\mathcal{M}$, that is, if it satisfies E1 and E2 relative to $\mathcal{K}$ in $\mathcal{M}$.*[29]

Explicitly Complete Explanation. *We will say that an explanation of ϕ relative to $\mathcal{K}$ in $\mathcal{M}$ is 'explicitly complete' if and only if it satisfies E1, E2 and $\vec{S}'=\varnothing$, that is: $\vec{X}=\vec{x}$ and $\vec{S}=\vec{s}$ are identical.*

Let us now see how these definitions deviate from the HP-definition. Most obviously, I have replaced the concept of a sufficient cause modulo $\mathcal{K}$ with the concept of a strong actual cause. But notice that I do not simply demand that the potential explanation $\vec{X}=\vec{x}$ is a strong actual cause of ϕ in every causal setting $\vec{u}\in\mathcal{K}$ with $(\mathcal{M},\vec{u})\models\vec{X}=\vec{x}\wedge\phi$. This would be the perfect analogue of condition EX1 in the HP-definition. Instead of demanding that $\vec{X}=\vec{x}$ itself must be a strong actual cause of ϕ, I demand that for being a potential or actual explanation $\vec{X}=\vec{x}$ must only be a part of an event $\vec{S}=\vec{s}$ that, according to $\mathcal{K}$, might be a strong actual cause of ϕ. The idea is basically this: An event that might be a strong actual cause of ϕ always amounts to an explicitly complete potential explanation of ϕ. But a potential or actual explanation does in general not have to be explicitly complete. It may very well leave out certain parts. Condition E2, which has no direct analogue in the HP-definition, regulates which parts of the complete causal explanation a potential or actual explanation may leave out:

[29] More correctly, one might say: $\vec{X}=\vec{x}$ is an explanation of ϕ relative to $\mathcal{K}$ in $\mathcal{M}$ if and only if it satisfies E1 E2, and E3 $\vee$ E4 relative to $\mathcal{K}$ in $\mathcal{M}$. But notice that E3 $\vee$ E4 is a tautology.

namely only information that the agent already fully believes. Notice next that the definitions of potential and actual explanation do not have an additional minimality condition, which rules out that explanations contain information that is already believed by the agent. So, in accordance with ERG, these definitions allow but do not demand the omission of informatively redundant parts. The definition of parsimonious potential explanation, on the other hand, demands the omission of informatively redundant parts (E5) in accordance with EPG. Notice, that in the HP-definition the minimality condition EX2 had the crucial job of precluding explanations with causally irrelevant conjuncts. In my definitions this job is already done by E1, since it includes the minimality condition SAC4 for the strong actual cause $\vec{S} = \vec{s}$. This is why the definitions of potential and actual explanation can do without an additional minimality condition like E5, without falling prey to counterexamples of causal irrelevance.

Next, notice that we do not need an analogue of condition EX3 from the HP-definition in our definitions. It already follows from condition E1 that the agent must consider $\vec{X} = \vec{x} \wedge \phi$ as possible. Also, while the HP-definition uses EX4 to separate trivial from nontrivial explanations, I use E3 (and E4 respectively) to differentiate between potential and actual explanations. I do not consider it useful to classify any potential explanation that an agent believes to be true as trivial. Just like it would be inapt to classify any knowledge of an agent as trivial, only because the agent already knows it. In Halpern's defence though, it makes sense to devaluate or even exclude explanations that are already believed by α if one aims for an explanation producing program (EPG style). The program does not need to provide α with information that she already has.[30] But for the recognition of explanations it makes no sense to devaluate $\vec{X} = \vec{x}$ as an explanation of a phenomenon ϕ for an agent α only because α already believes $\vec{X} = \vec{x}$ to be true. Instead, the belief in the truth of $\vec{X} = \vec{x}$ plus the fact that it is a part of an event that is known to be a strong actual cause of ϕ implies that α not only considers $\vec{X} = \vec{x}$ to potentially be an explanation of ϕ but that she takes it for an actual one.

Finally, the HP-definition demanded that the potential explanation $\vec{X} = \vec{x}$ must be a sufficient cause modulo $\mathcal{K}$ of ϕ in *every causal setting* $(\mathcal{M}, \vec{u})$ with $\vec{u} \in \mathcal{K}$ and $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{x} \wedge \phi$. Instead, I only demand that the potential explanation $\vec{X} = \vec{x}$ is part of an event $\vec{S} = \vec{s}$ that is a strong actual cause of ϕ in *at least one causal setting* $(\mathcal{M}, \vec{u})$ with $\vec{u} \in \mathcal{K}$. This adjustment takes into account our reformulation of EPG, according to which α must only believe that $\vec{S} = \vec{s}$ *might* be a strong actual cause of ϕ and not that it definitely is a strong actual cause of ϕ, given that it is the case. As pointed out above, I consider this weakening to be much more in accord with the concept of potential explanation. For a potential explanation, it is enough that it *might* amount to a strong actual cause of ϕ. It does not always have to amount to a strong actual cause whenever it is the case.[31]

I will now examine how the new definitions deal with the examples that proved to be problematic for the HP-definition.

[30] This is why there is no such thing as a parsimonious actual explanation.

[31] We will see in chapter 3, that this weakening is even essential for ensuring that our definition of potential explanation does not collapse into triviality in the context of probabilistic causal relationships.

1.5.4 The Overdetermined Forest Fire Revisited

According to the new definitions, $A = 1$ is no explanation of $F = 1$ relative to $\mathcal{K}^{OF}$ in $\mathcal{M}^{OF}$.

Proof. $A = 1 \wedge B = 1$ is the only possible choice for $\vec{S} = \vec{s}$, since it is the only possible strong actual cause of $F = 1$ of which $A = 1$ is a part. But with $\vec{S} = A \wedge B$, E2 is not satisfied because $(\mathcal{M}^{OF}, \vec{u}) \models \vec{S'} = \vec{s'}$ does not hold for all $\vec{u} \in \mathcal{K}^{OF}$, since $\vec{S'} = B$ and $\vec{s'} = 1$ and $(\mathcal{M}^{OF}, \vec{u_2}) \not\models B = 1$. □

According to the new definitions, $A = 1 \wedge B = 1$ is a parsimonious potential (and therefore a potential and no actual) explanation of $F = 1$ relative to $\mathcal{K}^{OF}$ in $\mathcal{M}^{OF}$.

Proof. Take $A = 1 \wedge B = 1$ itself for $\vec{S} = \vec{s}$. Then, there is a $\vec{u} \in \mathcal{K}^{OF}$, namely $\vec{u_0}$, such that $\vec{S} = \vec{s}$ is a strong actual cause of $F = 1$ in $(\mathcal{M}, \vec{u})$. Also, $A = 1 \wedge B = 1$ is clearly a part of $\vec{S} = \vec{s}$. So, E1 is satisfied. E2 is trivially satisfied, since $\vec{S'}$ is empty. E4 is satisfied, since, for example, $(\mathcal{M}^{OF}, \vec{u_2}) \models \neg(A = 1 \wedge B = 1 \rightarrowtail F = 1)$. As just shown, neither $A = 1$ satisfies E2 nor does $B = 1$ (the argument works analogously). E5 is therefore also satisfied. □

The new definitions therefore dismiss explanations which exclude causally relevant information only because α could in principle deduce the information by herself.

1.5.5 The Dry Forest Revisited

According to the new definitions, $D = 1$ is no explanation of $F = 1$ relative to $\mathcal{K}^D$ in $\mathcal{M}^D$.

Proof. It is easy to check that $D = 1$ alone is no strong actual cause of $F = 1$ in any setting for $\mathcal{M}^D$. So, take $D = 1 \wedge A = 1$ as the first candidate for $\vec{S} = \vec{s}$. Then E2 is not fulfilled because $(\mathcal{M}^{OF}, \vec{u_1}) \not\models A = 1$. The only alternative for $\vec{S} = \vec{s}$ is $D = 1 \wedge L = 1$. But here E2 is not fulfilled because $(\mathcal{M}^{OF}, \vec{u_0}) \not\models L = 1$. □

It is easy to check that $A = 1$ and $L = 1$ are (parsimonious) potential explanations of $F = 1$ relative to $\mathcal{K}^D$ in $\mathcal{M}^D$ according to the new definitions (and according to the HP-definition). These results are in accordance with the idea that, if α accepts an explanation, it should enable her to identify what she considers to be the causal process that is responsible for the realization of the explanandum. Further, $D = 1 \wedge A = 1$ and $D = 1 \wedge L = 1$ are potential, but no parsimonious potential explanations of $F = 1$ according to the new definitions.

1.5.6 Suzy, Billy, and the Bottle Revisited

According to the new definitions, $ST = 1 \wedge BT = 1$ is no explanation of $BB = 1$ relative to $\mathcal{K}^{SB}$ in $\mathcal{M}^{SB}$ because there is no strong actual cause $\vec{S} = \vec{s}$ of $BB = 1$ of which $ST = 1 \wedge BT = 1$ is a part. So, in contrast to the HP-definition, the new definitions do not allow explanations with causally irrelevant conjuncts. But just like the HP-definition, the new definitions recognize, for example, $SS = 1 \wedge ST = 1$ as a parsimonious potential explanation of $BB = 1$.

What about $BS = 1 \wedge BT = 1$? Take $BS = 1 \wedge BT = 1$ itself for $\vec{S} = \vec{s}$. Although $BS = 1 \wedge BT = 1$ is not a strong actual cause of $BB = 1$ in every causal setting, in which $BS = 1 \wedge BT = 1$ holds (see $(\mathcal{M}, \vec{u_1})$), it is still the case that $BS = 1 \wedge BT = 1$ is a strong actual cause of $BB = 1$ in some causal settings $(\mathcal{M}^{SB}, \vec{u})$ with $\vec{u} \in \mathcal{K}^{SB}$ (see $\vec{u_1}$, $\vec{u_3}$, and $\vec{u_4}$). Condition E1 is therefore satisfied. So is condition E2, since $\vec{S'} = \varnothing$. As just pointed out, $BS = 1 \wedge BT = 1$ is no strong actual cause of $BB = 1$ in $(\mathcal{M}, \vec{u_1})$, which means that E3 is also satisfied. So $BS = 1 \wedge BT = 1$ is acknowledged as a potential explanation of $BB = 1$ by our amended definition.

1.5.7 Strong Actual Causes vs. Sufficient Causes

At this point, I would like to consider a potential objection. As it turns out, there is an alternative to our amended definitions of explanation, that is able to resolve all the problematic examples that we have raised equally well. This alternative differs from our proposal in section 1.5.3 in one simple but crucial respect: Where we have employed the concept of strong actual causation, it uses the concept of sufficient causation:

Potential Explanation (Alternative Proposal). $\vec{X} = \vec{x}$ *is a 'potential explanation' of* ϕ *relative to a set* $\mathcal{K}$ *of contexts in a causal model* $\mathcal{M}$ *if and only if the following conditions hold:*

E1 There is an event $\vec{S} = \vec{s}$ *with* $\vec{X} = \vec{x}$ *being a part of* $\vec{S} = \vec{s}$ *and* $\vec{S} = \vec{s}$ *is a sufficient cause* ϕ *in* $(\mathcal{M}, \vec{u})$ *for some* $\vec{u}$ *in* $\mathcal{K}$*.*

E2 For all contexts $\vec{u} \in \mathcal{K} : (\mathcal{M}, \vec{u}) \models \vec{S'} = \vec{s'}$ *(where* $\vec{S'} = \vec{S} \setminus \vec{X}$ *and* $\vec{s'}$ *is the restriction of* $\vec{s}$ *to the variables in* $\vec{S'}$*).*

E3 $\vec{S} = \vec{s}$ *is no sufficient cause of* ϕ *in* $(\mathcal{M}, \vec{u'})$ *for some* $\vec{u'} \in \mathcal{K}_\phi$*.*

Actual Explanation (Alternative Proposal). $\vec{X} = \vec{x}$ *is an 'actual explanation' of* ϕ *relative to* $\mathcal{K}$ *in* $\mathcal{M}$ *if and only if it satisfies E1, E2, and:*

E4 $\vec{S} = \vec{s}$ *is a sufficient cause of* ϕ *in* $(\mathcal{M}, \vec{u'})$ *for all* $\vec{u'} \in \mathcal{K}_\phi$*.*

To see that these definitions yield just the same results for the presented examples as our amended definitions, notice that in all the examples that we have considered all strong actual causes of the given explanandum are also sufficient causes of the explanandum and vice versa.[32] One might therefore wonder: Do we really need the concept of strong actual causation? It seems that the adjustments that we have made in our amended definitions of explanation enable us to employ the concept of sufficient causation in these definitions, even though this concept seemed inadequate to explicate explanations in the setting of the original HP-definition. So, can we just achieve the same for our explications of *explanation* without reinvigorating the neglected concept of strong actual causation and simply use the much better established concept of sufficient causation?

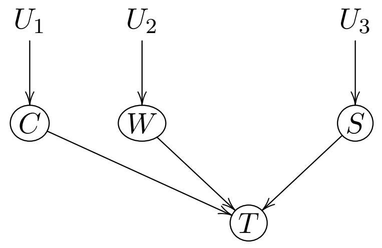
The answer is no! Even though the concept of sufficient causation performs equally well in our amended definitions of explanation with the considered problematic examples, we should refrain from using it. It is no adequate alternative to the concept of strong actual causation. But to see why, we have to wait until the end of chapter 2. By then we will see that using the concept of sufficient causation for explications of *explanation* would lead to some undesirable consequences.

[32] This is not always the case though. The concepts of strong actual causation and sufficient causation are not extensionally equivalent. Not even in scenarios in which all causal relationships are considered to be deterministic.

1.6 Partial and Ambivalent Partial Explanations

1.6.1 Partial Explanations and Explanatory Completeness

It is a crucial part of our definition of potential explanation that a full or complete potential explanation does not need to be explicitly complete. That means, an event $\vec{X} = \vec{x}$, that is only a smaller part of an explicitly complete potential explanation $\vec{X} = \vec{x} \wedge \vec{S'} = \vec{s'}$ of a given explanandum ϕ, will still be acknowledged as a full potential explanation of ϕ for an agent α, if α already fully beliefs that $\vec{S'} = \vec{s'}$ is the case. So far, we have pretended that this acknowledgement of non-explicitly complete potential explanations as potential explanations is an all or nothing matter: Given an explicitly complete potential explanation $\vec{X} = \vec{x} \wedge \vec{S'} = \vec{s'}$ of ϕ, $\vec{X} = \vec{x}$ is only recognized as a potential explanation of ϕ, if α fully beliefs $\vec{S'} = \vec{s'}$ ($\mathcal{P}(\vec{S'} = \vec{s'}) = 1$). But clearly, our intuitions concerning explanations do not follow such a black-and-white classification. Intuitively, there are various shades of grey. Consider the following example by Halpern and Pearl (2005b): We notice that our friend Victoria is unusually tanned. Since the weather at home could not have caused this tan, we seek for potential explanations. We quickly have two candidates at hand: Since Victoria's family owns a house in the Canary Islands, she might recently have taken a vacation there. Alternatively, Victoria might have become a regular customer of a tanning salon. These beliefs can be represented by the causal model $\mathcal{M}^V$ that is presented in figure 1.5 and in which C represents whether Victoria took a vacation in the Canary Islands, W represents whether it was sunny during this time in the Canary Islands, and S represents whether Victoria recently frequenced a tanning salon:



- $C := U_1$
- $W := U_2$
- $S := U_3$
- $T := (C \wedge W) \vee S$

Figure 1.5: $\mathcal{M}^V$ - tanned Victoria scenario.

We consider the following epistemic state $\mathcal{K}^V$, which includes the belief that Victoria is tanned ($T = 1$):

- $\mathcal{K}^V = \{\vec{u_0} = (1,1,1), \vec{u_1} = (1,1,0), \vec{u_2} = (0,1,1), \vec{u_3} = (1,0,1), \vec{u_4} = (0,0,1)\}$, with $\vec{u} = (u_1, u_2, u_3)$.

As already pointed out, an epistemic state $\mathcal{K}$ can be amended by a probability distribution $\mathcal{P}$ over the elements in $\mathcal{K}$ to represent an agents degrees of beliefs. So, to represent our rather strong belief that it has recently been sunny in the Canary Islands, we should use a probability distribution $\mathcal{P}$ that assigns rather low probabilities to $\vec{u_3}$ and $\vec{u_4}$.[33] Let us, as an example, consider the following distribution:

- $\mathcal{P}(\vec{u_0}) = 0.3$
- $\mathcal{P}(\vec{u_1}) = 0.3$
- $\mathcal{P}(\vec{u_2}) = 0.2$
- $\mathcal{P}(\vec{u_3}) = 0.1$
- $\mathcal{P}(\vec{u_4}) = 0.1$

Now, intuitively, we consider the information that Victoria recently took a vacation in the Canary Islands ($C = 1$) as a fairly good potential explanation of her tan ($T = 1$), even though $C = 1$ is not acknowledged as a potential explanation of $T = 1$ according to our definition, since $W = 1$, the information that it has recently been sunny in the Canary Islands, is not fully believed in the given epistemic state.[34] The reason that we intuitively still consider $C = 1$ as a fairly good potential explanation of $T = 1$ is that $W = 1$, though not being believed with perfect confidence, has a very high probability in our epistemic state. As soon as we lower the degree of belief in $W = 1$, for example to 0.5, $C = 1$ becomes a much worse potential explanation, worse, in the sense that its incompleteness becomes much more apparent. The higher the degree of belief in $W = 1$, the less apparent is the incompleteness of $C = 1$ as a potential explanation of $T = 1$. This is why $C = 1$ appears as a fairly good potential explanation of $T = 1$ relative to an epistemic state, in which $\mathcal{P}(W = 1)$ is high. But the lower the degree of belief in $W = 1$, the more apparent is the incompleteness of $C = 1$ as a potential explanation of $T = 1$. This motivates the following definition, which is inspired by Halpern and Pearl (2005b, p. 904), who explicate a concept of partial explanation in a similar way for their definition of explanation:

Partial Explanations and Degree of Explanatory Completeness. *Let $\mathcal{K}_{\vec{X}=\vec{x},\phi}$ be the largest subset of $\mathcal{K}$ such that $\vec{X} = \vec{x}$ is an explanation of ϕ relative to $\mathcal{K}_{\vec{X}=\vec{x},\phi}$ in $\mathcal{M}$ (which means: $\vec{X} = \vec{x}$ satisfies conditions E1, E2 relative to $\mathcal{K}_{\vec{X}=\vec{x},\phi}$ in $\mathcal{M}$). $\vec{X} = \vec{x}$ is a partial explanation of ϕ relative to $\mathcal{K}$ in $\mathcal{M}$ if and only if $\varnothing \neq \mathcal{K}_{\vec{X}=\vec{x},\phi} \subset \mathcal{K}$. The degree of explanatory completeness of the partial explanation $\vec{X} = \vec{x}$ is given by $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi})$.*

$\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi})$ is the probability (an agent's degree of belief) that $\vec{X} = \vec{x}$ amounts to a potential or actual explanation of ϕ. Clearly, we can apply this measure also to complete explanations, though the result will always be trivial: For any $\vec{X} = \vec{x}$ that is an explanation of ϕ relative to a given epistemic state $\mathcal{K}$, we have $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi}) = 1$, since $\mathcal{K}_{\vec{X}=\vec{x},\phi} = \mathcal{K}$. By demanding that for any partial explanation $\vec{X} = \vec{x}$ of ϕ, the set $\mathcal{K}_{\vec{X}=\vec{x},\phi}$ is a real subset of $\mathcal{K}$ we ensure that every partial explanation has a non-maximal degree of completeness: $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi}) < 1$. So the concepts of complete explanation and partial explanation do not intersect. Coming back to our example, the information that Victoria recently took a vacation in the Canary Islands ($C = 1$), is now acknowledged as a partial explanation of her unusual tan ($T = 1$) relative to $\mathcal{K}^V$ in $\mathcal{M}^V$, because $C = 1$ is a potential explanation of $T = 1$ relative to $\mathcal{K}^V_{C=1,T=1} = \{\vec{u_0}, \vec{u_1}, \vec{u_2}\}$ in $\mathcal{M}^V$. The degree of explanatory completeness of $C = 1$ is $\mathcal{P}(\{\vec{u_0}, \vec{u_1}, \vec{u_2}\}) = 0.8$.

Clearly, the degree of explanatory completeness is not the only measure of explanatory

[33] The degree of belief in an event like $\vec{X} = \vec{x}$ can be calculated as the sum of the degrees of belief in all those contexts $\vec{u}$ with $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{x}$.

[34] Notice that $C = 1 \wedge W = 1$ is an explicitly complete potential explanation of $T = 1$ in $\mathcal{M}^V$ relative to $\mathcal{P}$.

goodness or explanatory loveliness. Another natural measure of the loveliness of a potential explanation $\vec{X} = \vec{x}$ is its plausibility, which can be measured by $\mathcal{P}(\vec{X} = \vec{x})$. In chapter 8, we will discuss in some detail another dimension of explanatory loveliness, namely explanatory power, which aims to measure an explanation's ability to reduce an agent's surprise about the explanandum. So, the degree of explanatory completeness of an explanation should be seen as just one under several distinct dimensions of explanatory goodness.

As pointed out, my definition of partial explanation is a derivation from Halpern and Pearl's definition of partial explanation as given in (Halpern and Pearl, 2005b, p. 904). But I have made some crucial changes. First, Halpern and Pearl do not demand that the set $\mathcal{K}_{\vec{X}=\vec{x},\phi}$, relative to which the partial explanation $\vec{X} = \vec{x}$ is an explanation of ϕ, must be a real subset of $\mathcal{K}$. Instead it may also be identical to $\mathcal{K}$ itself. As a consequence, every complete explanation is also a partial explanation. This is a rather unintuitive utilization of the adjectives "partial" and "complete". More crucially, Halpern and Pearl employ a different measure of explanatory completeness.[35] Instead of $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi})$, the probability that $\vec{X} = \vec{x}$ is an explanation of ϕ, Halpern and Pearl propose to use $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi} \mid \vec{X} = \vec{x})$, the probability that $\vec{X} = \vec{x}$ is an explanation of ϕ, given $\vec{X} = \vec{x}$. But, given our new explication of explanation, $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi} \mid \vec{X} = \vec{x})$ is highly inadequate as a measure of explanatory completeness. The value of $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi})$ is given by the sum of the probabilities of those contexts, relative to which $\vec{X} = \vec{x}$ satisfies the conditions ($E1$, $E2$, $E3 \vee E4$) of being an explanation of ϕ. In the tanning-example above, $C = 1$ satisfies those conditions relative to the contexts $\vec{u_0}$, $\vec{u_1}$, and $\vec{u_2}$. It does not satisfy those conditions relative to $\vec{u_3}$ and $\vec{u_4}$, because $W = 1$, the remaining conjunct of the sufficient cause $C = 1 \wedge W = 1$ of which $C = 1$ is a part, does not hold in $\vec{u_3}$ and $\vec{u_4}$. So, $\vec{u_3}$ leads to a violation of condition E1, while both $\vec{u_3}$ and $\vec{u_4}$ lead to a violation of condition E2. This illustrates that not only contexts, in which the explanans $\vec{X} = \vec{x}$ does hold can lead to a violation of the conditions of being a potential explanation of ϕ. A context, in which $\vec{X} = \vec{x}$ does not hold, like $\vec{u_4}$ in our example, can still yield a violation of condition E2. Now, applying the measure $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi} \mid \vec{X} = \vec{x})$ would neglect such contexts. By conditionalizing on $\vec{X} = \vec{x}$, all contexts in which $\vec{X} = \vec{x}$ does not hold are disregarded. As a consequence, $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi} \mid \vec{X} = \vec{x})$ tends to overestimate the degree of explanatory completeness. To illustrate this, consider the following alternative epistemic state for our tanning example:

- $\mathcal{K}'^{V} = \{\vec{u_0} = (1,1,1), \vec{u_1} = (1,1,0), \vec{u_2} = (0,1,1), \vec{u_3} = (0,0,0), \vec{u_4} = (0,0,1)\}$,

with the following distribution:

- $\mathcal{P}(\vec{u_0}) = 0.2$
- $\mathcal{P}(\vec{u_1}) = 0.2$
- $\mathcal{P}(\vec{u_2}) = 0.2$
- $\mathcal{P}(\vec{u_3}) = 0.2$
- $\mathcal{P}(\vec{u_4}) = 0.2$

[35] Halpern and Pearl (2005b) do not use the term 'measure of explanatory completeness'. Instead they speak of a measure of 'explanatory goodness'. I prefer the term 'explanatory completeness' since this amounts to a clearer description of what is actually measured. I think that the term 'explanatory goodness' is better suited for a more general description of the value of an explanation. As pointed out in a previous paragraph, an explanation may be good or bad on several distinct dimensions. The degree of explanatory completeness is just one dimension of explanatory goodness, while explanatory power or the plausibility of explanations are other such dimensions.

$C = 1$ is again a partial explanation of $T = 1$ relative to $\mathcal{K}'^V$, while $C = 1 \wedge W = 1$ is a (complete) potential explanation of $T = 1$ relative to $\mathcal{K}'^V$. But when applying the measure $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi} \mid \vec{X} = \vec{x})$ as a measure of explanatory completeness, we get: $\mathcal{P}(\mathcal{K}_{C=1,T=1} \mid C = 1) = 1$ and $\mathcal{P}(\mathcal{K}_{C=1\wedge W=1,T=1} \mid C = 1 \wedge T = 1) = 1$. So, the partial explanation $C = 1$ of $T = 1$ is acknowledged to have the maximal degree of explanatory completeness and therefore the same degree of explanatory completeness as the complete potential explanation $A = 1 \wedge W = 1$. This is clearly a highly inadequate result for a measure of explanatory completeness. Using $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi})$ instead, we get: $\mathcal{P}(\mathcal{K}_{C=1,T=1}) = 0.6$ and $\mathcal{P}(\mathcal{K}_{C=1\wedge W=1,T=1}) = 1$, which means that, in accordance with intuition, the partial explanation $C = 1$ of $T = 1$ has a significantly lower degree of explanatory completeness than the complete potential explanation $C = 1 \wedge W = 1$ of $T = 1$.

With a definition of partial explanation at hand, let us now come back to one of our previously discussed examples. As it turns out, it provides an illustrative and informative application for our newly introduced concept of partial explanation.

1.6.2 Ambivalent Partial Explanations

Consider again the dry forest scenario. We have seen that, unlike the HP-definition of explanation, our amended definitions do not acknowledge $D = 1$ as a (complete) explanation of $F = 1$ relative to the epistemic state $\mathcal{K}^D$ in $\mathcal{M}^D$. I have argued that this is a reasonable result, because the information $D = 1$ does not provide α with a potential strong actual cause of $F = 1$. Still, $D = 1$ turns out to be a partial explanation of $F = 1$ relative to $\mathcal{K}^D$ according to the definition of partial explanation that we have introduced in the previous section. To see this, take $\vec{S} = D \wedge L$, then $D = 1$ is an explanation relative to $\{\vec{u_1}\}$ and therefore a partial explanation relative to $\mathcal{K}^D$ in $\mathcal{M}^D$ with a degree of explanatory completeness of $\mathcal{P}(\{\vec{u_1}\})$. But interestingly, $\{\vec{u_1}\}$ is not the only largest subset of $\mathcal{K}$, relative to which $D = 1$ is an explanation of $F = 1$ in $\mathcal{M}^D$. $D = 1$ is also an explanation relative to $\{\vec{u_2}\}$, where it is a part of the strong actual cause $D = 1 \wedge A = 1$. So, $D = 1$ is also a partial explanation of $F = 1$ relative to $\mathcal{K}^D$ in $\mathcal{M}^D$ with a degree of explanatory completeness of $\mathcal{P}(\{\vec{u_2}\})$. This motivates the following definition:

Ambivalent Partial Explanations . *If and only if there are at least two largest, non-empty, real subsets $\mathcal{K}_{\vec{X}=\vec{x},\phi}$, $\mathcal{K}'_{\vec{X}=\vec{x},\phi}$ of $\mathcal{K}$ such that $\vec{X} = \vec{x}$ is an explanation of ϕ relative to $\mathcal{K}_{\vec{X}=\vec{x},\phi}$ and relative to $\mathcal{K}'_{\vec{X}=\vec{x},\phi}$ in $\mathcal{M}$, then $\vec{X} = \vec{x}$ is an ambivalent partial explanation of ϕ relative to $\mathcal{K}$ in $\mathcal{M}$.*

The concept of an ambivalent partial explanation is a consequence of our amended definition of explanation, according to which it is a defining feature of an explanation to be a part of an event $\vec{S} = \vec{s}$ that might be a strong actual cause of the explanandum. Any partial or complete explanation of an explanandum ϕ can therefore be understood as a pointer to such a potential strong actual cause of ϕ. As soon as a piece of information is part of more than one potential strong actual cause of the explanandum, this piece of information is an ambiguous pointer. It is unable to identify a unique strong actual cause, that might have been responsible for the causal production of the explanandum. Since an ambivalent partial explanation is part of more than one complete explanation, it does not have a unique degree of completeness. Instead, its degree of completeness depends on the complete explanation that we compare it to. In our example, the degree of completeness of $D = 1$ is $\mathcal{P}(\{\vec{u_1}\})$ when compared to the complete explanation $D = 1 \wedge L = 1$. But when compared to the complete explanation $D = 1 \wedge A = 1$, then the degree of completeness of $D = 1$ is $\mathcal{P}(\{\vec{u_2}\})$.

Notice that the characterization of $D = 1$ as an ambivalent partial explanation of $F = 1$ captures exactly our concerns that we have expressed concerning the HP-definitions handling of this example, namely that $D = 1$ cannot be seen as a complete explanation of $F = 1$, because learning that $D = 1$ holds does not enable an agent with epistemic state $\mathcal{K}^D$ to identify the entire strong actual cause that is responsible for the production of $F = 1$. $D = 1$ does not function as an unambiguous pointer to an event that might be a strong actual cause of $F = 1$.

1.7 Correct Explanations

According to the HP-definition and the amended definitions as presented in section 1.5, whether something is an explanation of some phenomenon for an agent α is completely independent of what happens outside of α's head. It only hinges on α's beliefs about the world, while the world's actual constitution is irrelevant.[36] Craver (2007) argues that such an account misses something essential. He advances 'the idea that the norms of explanation fall out of a commitment by scientists to describe as accurately and completely as possible the relevant ontic structures in the world. Viewed in this way, our theories of scientific explanation cannot carve off those ontic structures as if they were expendable [...]' (Craver, 2007, p. 48). But this is exactly what the HP-approach does. It exclusively employs epistemic constraints that concern a given epistemic state $\mathcal{K}$ of some agent α, without taking into account what the external world really looks like. This is clearly a feature that lets the HP-approach stand out from other accounts of causal explanation, since most are in accord with Craver's demands by focusing on ontic constraints that refer to causal structures, which are supposed to describe as accurately as possible the external world.[37] Craver's view of the indispensability of ontic constraints is also endorsed by Illari (2013). Although Illari considers epistemic constraints to be just as important as ontic constraints, she agrees with Craver that ontic constraints are a must-have for any account of successful explanation.

The demand for ontic constraints is certainly not unreasonable. It conforms with the widespread idea that an explanation should be true or truthlike. But it places us in a dilemma. If ontic constraints are indispensable, the HP-approach, including my amendments of the HP-definition, has no chance of providing a legitimate account of explanation.

But there is a way to acknowledge a need for ontic constraints while still holding on to the HP-approach. I agree with Craver and Illari that ontic constraints are indeed essential, but only for a certain conception of explanation which I will call *correct explanation.* This

[36] The constraints in the definitions refer not only to an epistemic state $\mathcal{K}$, but also to a causal model $\mathcal{M}$. But notice that $\mathcal{M}$ is understood to represent α's beliefs about the causal laws that govern a given scenario. So far, there is no condition that requires those beliefs to be true.

[37] See, for example, (Woodward, 2003), (Strevens, 2008), or (Machamer et al., 2000). Strevens explicitly states: "I follow the lead of most philosophers of explanation, and of most proponents of the causal approach in particular, in giving the [...] ontological sense of explanation precedence" (Strevens, 2008, p. 6). He describes the ontological sense of explanation as a conception according to which an explanation is 'something out in the world, a set of facts to be discovered' (Strevens, 2008, p. 6).

conception requires the identification of the causal process that is actually responsible for the production of the explanandum. But with the HP-approach comes the claim that there are other meaningful conceptions of explanation besides correct explanations, namely those that allow for an explanation to be incorrect. Potential, parsimonious potential, actual, and explicitly complete explanations as defined above are four examples. For such concepts the use of ontic constraints is completely unwarranted. To see that concepts like these are not futile, we only have to point to projects like the explication of Inference to the Best Explanation,[38] where an evaluation of potential (possibly incorrect) explanations from the point of view of a certain epistemic state is needed.

But although the HP-approach as presented in (Halpern and Pearl, 2005b) and (Halpern, 2016) focuses on the definition of epistemically relative explanations, this does not mean that it is unable to explicate the concept of correct explanations.[39] We can easily amend our definition of actual explanation with an ontic constraint to obtain a definition of correct explanation:

Correct Explanation. $\vec{X} = \vec{x}$ *is a 'correct explanation' of* ϕ *relative to an epistemic state* $\mathcal{K}$ *in a causal model* $\mathcal{M}$ *if and only if* $\vec{X} = \vec{x}$ *is an actual explanation of* ϕ *relative to* $\mathcal{K}$ *in* $\mathcal{M}$ *and* $\mathcal{K}$ *does not entail any false beliefs and* $\mathcal{M}$ *is an accurate description of the ontic causal structure of the world.*[40]

Notice that according to this definition correct explanations are, just like potential and actual explanations, agent-relative, which means that every correct explanation is a correct explanation for some agent in a certain epistemic state. It is of course possible to define an agent-independent concept of correct explanation, for example, by stating that a correct explanation of a phenomenon ϕ simply is a strong actual cause of ϕ in an empirically adequate causal model $\mathcal{M}$. Here, being a correct explanation is relative to a causal model, but not relative to any agent or epistemic state, which, incidentally, also makes the definition of correct explanation less complex. But empirical results by Waskan et al. (2014), as well as Wilkenfeld and Lombrozo (2020), suggest that laypeople, as well as scientists, consider the generation of understanding for some agent as constitutive for explanations. Since understanding is something that happens to an agent in a certain epistemic state, these results clearly speak against an agent-independent conception of explanation and give further justification for the HP-approach that emphasizes the essentiality of epistemic constraints.

1.8 Summary

Although the HP-definition constitutes a very promising approach to a strictly formalized explication of causal explanation, it does not live up to its own demands that are summarized in EPG. As a consequence, the HP-definition delivers problematic results in several examples, even if we presuppose that the concepts of actual and sufficient causation are adequately defined. But

[38] See, for example, (Aliseda, 2000).

[39] Halpern and Pearl (2005b, p. 901) already point out that it is possible to define something like correct explanations within their approach, although they abstain from doing so.

[40] To render this constraint usable, we have to explicate how to determine if a causal model accurately describes the ontic causal structure of the world, which is done by a methodology of causal induction as described by (Pearl, 2000) or (Spirtes et al., 2000).

the shortcomings of the HP-definition can be overcome while retaining its promising features. To achieve this, I have proposed several amendments to the HP-definition. The resulting account of causal explanation encompasses several distinct, but related concepts of causal explanation. The definitions of explicitly complete, potential, and actual explanation are in accordance with ERG and therefore ideal for the recognition of explanations. The definition of parsimonious potential explanation is in accordance with EPG and therefore ideal for producing causal explanations. I have also adapted the explication of partial explanations and the corresponding measure of explanatory completeness. All proposed definitions remedy the identified deficiencies of the HP-definition.

Further, the definition of actual explanation can be supplemented with an ontic constraint to explicate the notion of correct explanation. But just like the other concepts of explanation, correct explanations are considered to be agent-dependent, which is a supposition that lets the HP-approach stand out from other accounts of causal explanation and makes it not only a promising descriptive model of causal explanation, but also a valuable tool for projects like the formalization of Inference to the Best Explanation.[41]

[41] Chapter 1 is an adapted and expanded version of (Borner, forthcoming), which has been accepted for publication by *The British Journal for the Philosophy of Science* on August 08, 2021.

Chapter 2

Actual Causation in Probabilistic Contexts

2.1 Admitting Uncertainties

In the previous chapter I have only considered situations with completely deterministic causal relationships. In such cases, the value of an endogenous variable is completely determined by the values of its parents without the leeway for any influx of chance. This is clearly a significant restriction. Very often, causal relationships between events (or event-types) can only be described probabilistically.[1] Consider as an example the effectiveness of a certain drug. In epidemiology one typically uses randomized controlled trials (RCTs) to test and measure, whether and how well a drug D heals a certain symptom E. Even if it is significantly more effective than a placebo, a drug is typically not able to eliminate the symptom with perfect reliability. In most cases, the drug is only able to eliminate the symptom in a certain (non-maximal) percentage of those patients who take the drug. So, whenever an individual patient takes D to resolve the symptom E, there remains a non-negligible uncertainty, whether the comsumption of D actually leads to the desired effect in the given situation.

Alternatively, consider Suzy's stone-throwing abilities. It is completely unrealistic, however sober she may be, that Suzy is such an impeccable stone-thrower that there is not even the slightest chance of her missing the bottle. Since we are dealing with a human being instead of an infallible stone-throwing-machine, every realistic representation of the situation should therefore take the chance of missing into account. So, here again, when Suzy throws a stone, it will not hit the bottle with complete certainty. It will only do so with a certain (non-maximal) probability.

2.1.1 Probabilistic Causal Relationships and (In-)Determinism

One might want to argue that both examples that I have just given are not really cases of genuine probabilistic causal relationships and that genuine probabilistic causal relationships, if they exist at all, are rather unusual. In both examples, it may well be the case that we have given a very

[1] As we will later see, it is still up to debate, whether a probabilistic causal relationship can only exist between event-types or also between token events.

imprecise or even incomplete description of the cause and if we only describe the cause more thoroughly, we will see that there is a deterministic causal relationship between cause and effect after all. In case of the drug D, there might be some preconditions C in a patients organism that are sufficient for enabling D to resolve symptom E. So, when we identify those preconditions and consider them as being part of the cause, we will see that the causal relationship between $C \wedge D$ and the effect (the healing of E) is perfectly reliable and therefore deterministic. Now consider Suzy's stone throw. As soon as we give a specific description of the stone's mass and form, of the momentum that Suzy transmits to the stone and the atmospheric conditions, we will see that the relationship between Suzy's throw and the stone's collision with the bottle is also completely determinstic. Genuine probabilistic causal relationships only exist in genuinely indeterministic systems, which, if they exist at all, seem to be resticted to certain domains and phenomena of fundamental physics, like the behaviour of quantum particles or the decay of nuclei. If this is true, then probabilistic causal relationships are only a marginal phenomenon, restricted to events that have non-trivial objective chances, and deterministic causal models, as employed in the previous chapter, will clearly suffice, if not for all, then for most real-life causal scenarios.

I consider this view to be misguided. Even if we say that a causal relationship between events is only genuinely probabilistic if the probabilities of the events involved are objective chances, this does not necessarily mean that probabilistic causal relationships are rare and restricted to some domains in fundamental physics. Authors like Glynn (2010) or List and Pivato (2015) have convincingly argued that objective chances are level-relative and can therefore exist in higher level systems (or descriptions) of the world, even if the systems on which these higher level systems supervene are deterministic.[2] If this is correct, it means that genuinely probabilistic causal relationships (probabilistic causal relationships between events that have non-trivial objective chances) can emerge anywhere: between physical events just as well as between biological, psychological, economic, or social events.

But even if we do not belief in the existence of higher level objective chances, even if we think that our world is deterministic through and through (on every level) and non-trivial probabilties can only be epistemic probabilities that express our own uncertainties about the world, describing causal relationships probabilistically is still reasonable, even essential, for causal reasoning. Due to our epistemic and practical limitations, a precise and complete description of causal relations in a deterministic system is in many cases simply not feasible. Resorting to more abstract and partial descriptions and thereby introducing uncertainties into the desciption of causal relations is therefore often the best or even the only viable option.

Clearly, in this dissertation I do not want to solve the row between determinism and indeterminism. Neither do I want to pick a side. Instead, when I use the term 'probabilistic causal relationship' I will understand it to encompass both, probabilistic causal relationships that arise due to genuinely indeterministic systems (if those do indeed exist) and causal relationships that turn out to be probabilistic because of imprecise or incomplete descriptions of deterministic causal relationships. The accounts of probabilistic causal relationships that I will discuss in this

[2] For a detailed discussion and explication of the concept of levels see (List, 2019). In chapter 9, I will also give a more detailed description of levels.

chapter, will ultimately be compatible with both determinism and indeterminism, even though their respective proponents sometimes seem to suggest otherwise.

2.1.2 De Facto Dependence vs. De Facto Probability Raising

Another issue that does not coincide with the row between determinism and indeterminism, is the question of whether probabilistic causal relationships can hold between token events or whether they can only emerge on the type-level. The answer depends on what exactly we understand actual (token) causation to be in the context of probabilistic causal relationships. First of all, it is rather uncontroversial that probabilistic (as well as deterministic) causal relationships can typically not be observed on the token level. Instead, we typically use relative frequency data, ideally obtained from RCTs or at least from large data sets in which we are able to control for confounding factors, to deduce whether there is a (probabilistic) causal relationship between two factors or event-types and if there is, how strong this causal relationship is.[3] Now imagine that a probabilistic causal relationsip between two repeatable event-types C and E has been established. We have seen that in 90% of all cases, in which a C-type event was realized by an intervention, an E-type event happened, even though no other potential cause of an E-type event was present. There are at least two different accounts available to grasp what exactly happens on the token level. According to one view, we might say that every token of a C-type event produces a certain probability, like a single-case objective chance, of an E-token. This probability is typically higher than the probability of the E-token before C happened. Imagine, for example, that every C-token produces a probability of 0.9 of an E-token in circumstances in which no other potential causes of E are present. When we now look at, say, circa 10.000 repetitions of C-type events, we will notice that circa 9000 E-type events take place, because in every single instance, a C-token yielded a probability of 0.9 of an E-token. So, we have 10.000 potential E-tokens, each of which has a probability of 0.9 to happen. According to the law of large numbers, the number of E-tokens that actually happen is probably around 9000. Notice that what grounds the probabilistic causal relationship between C-type and E-type events is the production of a single-case probability of an E-token by every single C-token. Every C-token is therefore a probability raiser of an E-token. This is why it makes sense to say, that the causal relationship between a given C-token and an E-token is inherently probabilistic, since it is a (non-trivial) probability of the E-token that the C-token produces. I will call this account of actual causation, the *de facto probability raising* account of actual causation.[4]

Now, here is an alternative picture of what happens at the token level, when we have a probabilistic causal relationship between C-type and E-type events. Imagine again that a probabilistic causal relationsip between C and E has been established and we have seen that in 90% of all cases, in which a C-type event was realized by an intervention, an E-type event happened, even though no other potential cause of an E-type event was present. Now, instead of claiming that every single C-token produced a probability of 0.9 of a potential E-token, we might also say that

[3] In chapter 4, I will explore this process of causal induction in more detail. We will especially deal with the question of how to determine the strength of a probabilistic causal relationship based on relative frequency data. In the present chapter, we will mostly presuppose that the existence of a probabilistic causal relationship between two event-types, including its strength, is already established.

[4] We will later see, why the addition of 'de facto' is relevant.

90% of all C-tokens produced an E-token, while 10% did not.[5] Here it is not the production of a certain probability value on the token-level that grounds the probabilistic causal relationship on the type-level. Instead, it is the production of token events. For a specific C-token, it does not make any sense to speak of a probabilistic causal relationship to an E-token, because the C-token either does produce an E-token or it does not. The token causal relationship itself is therefore not intrinsically probabilistic. Still, probabilities come into play on a type-level and on an epistemic level. Although for any single C-token, it either does produce an E-token or it does not, it holds for a large number of C-tokens that only 90% of them produce E-tokens. And whenever we are incapable to determine which C-tokens will produce E-tokens and which will not (this incapability may be due to our epistemic or practical limitations or due to a genuinely indeterministic system), there will be an uncertainty about whether a given C-token will ultimately produce an E-token or not. I will call this account of actual causation, the *de facto dependence* account of actual causation, since in every single instance, an E-token either is produced by a C-token and therefore de facto depends on it, or it is not.

We can now see that the question of whether token events can stand in a probabilistic causal relationship to each other is connected to the issue of what grounds a probabilistic causal relationship on the type-level. Is it the production of the token effect by the token cause or is it the production of a certain probability of the token-effect by the token cause? Or to put it differently: Is actual causation the de facto dependence of one event on another event or is it the de facto dependence of one events probability on another event?

In the upcoming sections, I will deal more thoroughly with these questions. I will examine two different proposals from the literature to define actual causation in contexts of probabilistic causal relationships. Both proposals are based on two different, though related, causal model frameworks. I will show that both proposals, even though they may first appear equivalent, actually lead to different concepts of actual causation. I will also show that this divergence is essentially grounded in the underlying formal framework that each account employs. I will also show that this divergence follows exactly the differentiation between actual causation as de facto event-dependence and de facto probability raising.

The first task, we now have to take on is this: We have to tweak the framework of causal models, as it was introduced in the previous section, to allow for probabilistic causal relationships. The following, rather straightforward, idea can serve as a vantage point for this project: To incorporate probabilistic causal relationships into causal models we can replace the deterministic structural equations that assign definite values to each endogenous variable in dependence of the values of its parents with a function that assigns probability distributions over the values of each endogenous variable in dependence of the values of its parents. Take for example the structural equation in $\mathcal{M}^{SB}$ that connects the variable SH (whether Suzy hits) with the variables ST (whether Suzy throws) and SS (whether Suzy is sober) in the following way: $SH := (ST \wedge SS)$. This deterministic structural equation tells us that the event *Suzy hits* will definitely be realized when the events *Suzy throws* and *Suzy is sober* are realized. A probabilistic analogue to the deterministic structural equation should instead return a probability distribution over the values

[5] I will later deal in more detail with the concept of causal production. The use of the concept here is somewhat superficial, since, as it will later turn out, not every actual cause causally produces its effect in the exact meaning of the word. It would be more accurate to say: It influences the causal production of the effect.

of SH when receiving the values of ST and SS as arguments. Let us now see, how this idea can be worked out precisely.

2.2 Causal Bayes Nets as Probabilistic Causal Models

2.2.1 Introducing Causal Bayes Nets

A well entrenched way to explicate the idea of probabilistic causal models is the framework of Bayesian Networks.[6] Bayesian Networks are basically a tool to simplify the representation of and the calculation with a given probability distribution. To illustrate why, consider a set of variables $\mathcal{V} = \{V_1, ..., V_n\}$ with respective value sets $\mathcal{R}(V_1)$, ..., $\mathcal{R}(V_n)$ and a probability distribution $\mathcal{P}$ over a σ-algebra on the cartesian product of the value sets $\mathcal{R}(V_1)$, ..., $\mathcal{R}(V_n)$. Let $(v_1, v_2, ..., v_n) \in \mathcal{R}(V_1) \times ... \times \mathcal{R}(V_n)$ be a setting of the variables in $\mathcal{V}$. According to the chain rule, which simply follows from a repeated application of the definition of conditional probability, the probability $\mathcal{P}(v_1, v_2, ..., v_n)$ can be represented as follows:

$$\mathcal{P}(v_1, v_2, ..., v_n) = \mathcal{P}(v_1)\mathcal{P}(v_2|v_1)...\mathcal{P}(v_n|v_1, v_2, ..., v_{n-1}) = \prod_{i=1}^{n} \mathcal{P}(v_i|v_1, ..., v_{i-1}) \tag{2.1}$$

Now, consider a directed acylic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the elements in $\mathcal{V}$ as vertices and the elements of $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ being directed edges between two vertices. If there is an edge from V_1 to V_2 ($V_1 \rightarrow V_2$) in $\mathcal{E}$, we will say that V_1 is a *parent* of V_2 and V_2 is a *child* of V_1. The set of all parents of a variable V will be denoted by PAR_V, while par_V denotes a tuple of values for the variables PAR_V. We call a chain of directed edges that all point into the same direction a *directed path.* If there is a directed path from V_1 to V_2, we call V_1 an *ancestor* of V_2 and V_2 a *descendant* of V_1. ANC_V is the set of all ancestors of the variable V and DES_V is the set of all descendants of V. The tuple $(\mathcal{G}, \mathcal{P})$, consisting of the DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and the probability distribution $\mathcal{P}$ over the cartesian product of the value sets of the variables in $\mathcal{V}$, is a *Bayesian Network* if and only if $\mathcal{P}$ fulfills the Markov Condition relative to $\mathcal{G}$, which is:[7]

Markov Condition. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a DAG. $\mathcal{P}$ fulfills the Markov Condition relative to $\mathcal{G}$ if and only if, conditional on its parents in $\mathcal{G}$ every variable in $\mathcal{V}$ is, according to $\mathcal{P}$, probabilistically independent of its non-descendants in $\mathcal{G}$.*

If the Markov Condition is fulfilled, then the DAG $\mathcal{G}$ entails information about conditional independencies in $\mathcal{P}$, which can be used for a simpler calculation of the values of the joint probability distribution $\mathcal{P}(v_1, v_2, ..., v_n)$ by using the following truncated factorization:

$$\mathcal{P}(v_1, v_2, ..., v_n) = \prod_{i=1}^{n} \mathcal{P}(v_i|v_1, ..., v_{i-1}) = \prod_{i=1}^{n} \mathcal{P}(v_i|\ par_{V_i}) \tag{2.2}$$

[6] For a more thorough introduction into Causal Bayes Nets, see (Pearl, 2000) or (Williamson, 2004), whose expositions of Causal Bayes Nets I largely follow in this section.

[7] For a more thorough discussion of the Markov Condition and its consequences, see, for example, (Pearl, 2000, p. 14 ff.).

So far, we have treated Bayesian Networks as a tool for representing probabilistic dependencies and conditional independencies. But another interpretation is possible, one that augments the probabilistic relationships in a Bayesian Network with causal meaning. In this causal interpretation, any directed edge from V_1 to V_2 is understood to represent a direct causal influence from V_1 on V_2 and it is this causal influence that induces the probabilistic dependence that is found in the distribution $\mathcal{P}$. This causal interpretation of Bayesian Networks is motivated by the well entrenched idea that probabilistic and causal dependencies are deeply connected. This supposed connection is famously expressed in Hans Reichenbach's Principle of Common Cause, which basically says that there is no probabilistic correlation without a causal grounding:

Principle of Common Cause (PCC). *If two events $V_1 = v_1$ and $V_2 = v_2$ are probabilistically dependent, then either $V_1 = v_1$ is a cause of $V_2 = v_2$, $V_2 = v_2$ is a cause of $V_1 = v_1$, or $V_1 = v_1$ and $V_2 = v_2$ are effects of common causes U, such that $V_1 = v_1$ and $V_2 = v_2$ are probabilistically independent, given U.*

In a causal interpretation of a Bayesian Network, the Markov Condition also adopts a causal meaning, which is captured by the Causal Markov Condition (CMC):

Causal Markov Condition (CMC). *Conditional on its direct causes, any variable in a Bayesian Network is probabilistically independent of its non-descendants.*

Now, the causal interpretation of Bayesian Networks can be justified by the fact that the CMC implies the PCC, which means that any causally interpreted Bayesian Network connects causal influence with probabilistic dependence in just the way, the PCC demands.[8]

But it does not follow from the CMC that every causally interpreted Bayesian Network is causally adequate in the way that it represents the real causal connections in a given scenario. This becomes clear, when we consider that for a given probability distribution $\mathcal{P}$ there can, in general, be more than one DAG relative to which $\mathcal{P}$ fulfills the Markov Condition. As Forster et al. (2018) point out: "Though the CMC tells us that some arrows must be included in a given graph, it never tells us which arrows should not be included in a given graph" (Forster et al., 2018, p. 825). This boils down to the fact that the probabilistic dependencies in a given distribution can be explained by several different causal models. So, how can we make sure that we find just the right Bayesian Network for a given probability distribution $\mathcal{P}$, which, if interpreted causally, is the correct causal model for the given scenario? This challenge is nothing else than the problem of causal induction. Pearl (2000) and Spirtes et al. (2000) have presented algorithms to find an adequate Causal Bayes Net for an empirically given probability distribution. It has been shown that we need to impose further condtions like Faithfulness, Frugality or some kind of Minimality to render the process of causal induction more effective.[9] I will not go into the problem of causal induction here. Instead I will simply presuppose that any Causal Bayes Net that I will consider in the following is such, that a given agent α considers it to be an adequate representation of the real causal relationships underlying the represented

[8] See (Williamson, 2004) for a proof of this claim and for an in depth discussion of the Principle of Common Cause.

[9] See, for example, (Forster et al., 2018) for a discussion of Faithfulness, Frugality, or Minimality in the context of causal induction.

scenario. But it should be pointed out that, although Faithfulness, Frugality or Minimality are useful conditions for causal induction, not every causally adequate Causal Bayes Net is indeed faithful, maximal frugal or minimal.[10]

2.2.2 Causal Bayes Nets as Causal Models

Having introduced the basic framework of Causal Bayes Nets, we can now examine if they actually meet the requirements that we have established for a probabilistic version of causal models. We are looking for a replacement of the deterministic structural equations as found in structural equation models: Instead of assigning each variable a definite value in dependence of the values of its parents, we want an assignment of a probability distribution over the values of a variable in dependence of the values of its parents. As it turns out, this is exactly what a Causal Bayes Net (CBN) gives us.

To illustrate, let us consider a Causal Bayes Net $(\mathcal{G}, \mathcal{P})$ that consists of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a probability distribution $\mathcal{P}$ over the cartesian product of the value sets of the variables in $\mathcal{V}$. The probabilistic analogue of deterministic structural equations is embedded in the probability distribution $\mathcal{P}$. $\mathcal{P}$ does not only entail information about the probability of a specific variable X taking on some value x, it also entails the probability of the event $X = x$, conditional on other events that are represented in the CBN. Especially, it can tell us the probability of $X = x$, given that the parents PAR_X of X in $\mathcal{G}$ take on certain values par_X. When we consider $\mathcal{P}(X = x_i | PAR_X = par_X)$ for every possible value x_i of X, then we get the causal information we are looking for: the probability distribution over the values of X in dependence of the values of X's parents. We can highlight these causally relevant information that are embedded in the probability distribution $\mathcal{P}$ by defining a Causal Bayes Net in reference to a so called probability specification $\mathcal{S}$. Williamson (2004) defines a probability specification $\mathcal{S}$ in a CBN $(\mathcal{V}, \mathcal{E}, \mathcal{P})$ as follows: For each variable $V \in \mathcal{V}$, $\mathcal{S}$ specifies the probability distribution over the value set of V conditional on the values of V's parents (Williamson, 2004, cf. p. 14). So, for each variable $V \in \mathcal{V}$, $\mathcal{S}$ provides for each value v of V and for each combination of values par_V of PAR_V a statement of the form: $\mathcal{P}(V = v | PAR_V = par_V) = p$. Williamson then shows that a probability specification $\mathcal{S}$ for a DAG $\mathcal{G}$ determines a unique probability distribution $\mathcal{P}$ over the cartesian product of the value sets of the variables in $\mathcal{G}$ (Williamson, 2004, cf. p. 17). A probability specification $\mathcal{S}$ is therefore just another way to describe a probability distribution $\mathcal{P}$ in a Causal Bayes Net $(\mathcal{G}, \mathcal{P})$. What makes a probability specification so useful for our purposes, is that the conditional probabilities, like $\mathcal{P}(V = v | PAR_V = par_V)$, which embody important causal information, are directly given. This is why, in the following, I will typically describe the probability distribution $\mathcal{P}$ of a Causal Bayes Net $(\mathcal{G}, \mathcal{P})$ by presenting the respective specification $\mathcal{S}$ that encodes $\mathcal{P}$.

As an example, consider a probabilistic version of the forest fire scenario, in which a lightning strike as well as a dropped match by an arsonist can cause a forest fire. Thanks to a great collection of data, we know that an arsonist dropping a match $(A = 1)$ yields a forest fire with a probability of 0.7, if no other causal influence on variable F is present. Independently from the arsonist, a lightning strike $(L = 1)$ also yields a forest fire $(F = 1)$ with a probability of

[10] For examples see (Forster et al., 2018).

0.7, if no other causal influence on variable F is present. We further know that the probability of an arsonist dropping a match in the forest is 0.05 in the given situation and the probability of a lightning strike hitting a tree in the forest is 0.1. The situation can be represented by the Causal Bayes Net $(\mathcal{G}, \mathcal{P})$, with $\mathcal{G}$ looking like this:

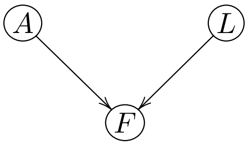


Figure 2.1: DAG for the probabilistic forest fire scenario.

The probability distribution $\mathcal{P}$ is encoded by the following specification $\mathcal{S}$:

- $\mathcal{P}(A = 1) = 0.05$
- $\mathcal{P}(L = 1) = 0.1$
- $\mathcal{P}(F = 1|A = 1, L = 1) = 0.91$
- $\mathcal{P}(F = 1|A = 1, L = 0) = 0.7$
- $\mathcal{P}(F = 1|A = 0, L = 1) = 0.7$
- $\mathcal{P}(F = 1|A = 0, L = 0) = 0$

The specification $\mathcal{S}$ directly provides us with all the important causal information that is embedded in the distribution $\mathcal{P}$. The last four conditional probabilities, which describe the probabilities of all values of F in dependence of all possible values of F's parents, can be seen as the probabilistic analogue of the structural equation with the variable F on the left-hand side.

Notice that in a CBN, just like in a structural equation model (SEM), we can differentiate between endogenous and exogenous variables, even though this distinction is typically not explicitly made in standard presentations of CBN's. In the example above, A and L are both exogenous, because they do not have any parents in the model. So, in CBNs the exogenous variables simply form a subclass of the variable set $\mathcal{V}$.[11] The specification $\mathcal{S}$ entails statements of the form '$\mathcal{P}(V = v|PAR_A = par_A) = p$' for all variables $V \in \mathcal{V}$, including the exogenous variables for which the set PAR_V is empty. For these variables the specifier simplifies to $\mathcal{P}(V = v) = p$ for each value v of V. The specifiers for the exogenous variables can be seen as the analogue to the context in an SEM. While they do not necessarily determine a unique value for each exogenous variable, they do provide us with a probability distribution over the values of each exogenous variable. This points to a crucial difference between the formalism of deterministic SEM's, as presented in chapter 1, and CBN's: In the framework of deterministic SEMs we differentiated between a causal model $\mathcal{M}$ and a causal setting $(\mathcal{M}, \vec{u})$. The causal model $\mathcal{M}$ entails all the information about the causal relationships in the represented scenario, but it does not contain any information about the actual values of the contained variables. This is the information that

[11] Remember that in SEMs the exogenous and the endogenous variables are formally disambiguated by introducing two distinct sets of variables: $\mathcal{V}$ and $\mathcal{U}$. This is typically not done in CBNs.

comes with the context $\vec{u}$. In a CBN $(\mathcal{G}, \mathcal{P})$, on the other hand, the probability distribution $\mathcal{P}$ provides us with both: quantitative information about the causal relationships in the represented scenario and information about the probabilities of the values of all the variables in the model.

Using a CBN as a causal model we can now define a formal language in analogy to the language that we have introduced for deterministic causal models in chapter 1. 'Primitive events' are now formulas of the form '$P(\vec{X} = \vec{x}) = p$' (read as: the probability of the event $\vec{X} = \vec{x}$ is p). Notice that 'P' is part of the object language, in contrast to '$\mathcal{P}$', which is part of the meta-language. We will say $(\mathcal{G}, \mathcal{P}) \models P(\vec{X} = \vec{x}) = p$ if and only if $\mathcal{P}(\vec{X} = \vec{x}) = p$ in $(\mathcal{G}, \mathcal{P})$.

Of course, we do not only want to be able to evaluate statements like $P(\vec{X} = \vec{x}) = p$ in probabilistic causal models. For defining causal concepts in probabilistic causal models, we also need the probabilistic analogue of intervention counterfactuals, which are statements of the form: 'The probability of the event $Y = y$ would be p, if the event $\vec{X} = \vec{x}$ would be brought about by an intervention.' But for being able to evaluate such statements in CBN's, we first need to introduce the concept of an intervention into the formal framework of CBN's.

2.2.3 The do-operator in Causal Bayes Nets

We can introduce interventions, which enable to set certain variables $\vec{X}$ to certain values $\vec{x}$, into CBNs by defining a *do*-operator $do(\vec{X} = \vec{x})$ that transforms a probability distribution $\mathcal{P}$ of a Causal Bayes Net $(\mathcal{G}, \mathcal{P})$ into a new probability distribution $\mathcal{P}_{do(\vec{X}=\vec{x})}$.[12] Since a probability distribution in a Causal Bayes Net is completely determined by its specification, we can define the operator $do(\vec{X} = \vec{x})$ by characterizing how it transforms the specification of $\mathcal{P}$. Consider an arbitrary specifier '$\mathcal{P}(Y = y|PAR_Y = par_y) = p$' for a probability distribution $\mathcal{P}$ of a CBN $(\mathcal{G}, \mathcal{P})$. We define $\mathcal{P}_{do(\vec{X}=\vec{x})}(Y = y|PAR_Y = par_y)$ as:[13]

$$\mathcal{P}_{do(\vec{X}=\vec{x})}(Y = y|PAR_Y = par_y) = \begin{cases} \mathcal{P}(Y = y|PAR_Y = par_y) & \text{if } Y \notin \vec{X} \\ 1 & \text{if } Y \in \vec{X} \text{ and } y \in \vec{x} \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

As the second and third case of the definition show, an intervention on a variable Y makes the probability distribution over the values of Y completely independent from its parents.[14] This is consistent with the workings of interventions in structural equation models. For deterministic structural equations we further supposed that a structural equation is invariant under interven-

[12] In the literature, the notation $\mathcal{P}(\phi \mid do(\vec{X} = \vec{x}))$ is often used as an alternative to $\mathcal{P}_{do(\vec{X}=\vec{x})}(\phi)$. This notation reminds of the conditional probability $\mathcal{P}(\phi \mid \vec{X} = \vec{x})$. But $\mathcal{P}_{do(\vec{X}=\vec{x})}(\phi)$ has a very different meaning than $\mathcal{P}(\phi \mid \vec{X} = \vec{x})$. The former gives us the new probability of ϕ after a change in the real world, which has been induced by an action that brings about the event $\vec{X} = \vec{x}$. The latter gives us the new probability of ϕ after a change of information that has been induced by learning the previously unknown information that $\vec{X} = \vec{x}$. Pearl puts it like this: "conditioning represents passive observations in an unchanging world, whereas actions change the world" (Pearl, 2000, p. 110).

[13] See (Kinney, 2019b), where the *do*-operator is defined likewise as an operation on the probability distribution in a CBN.

[14] Notice that in Causal Bayes Nets it is formally possible to intervene on exogenous variables. The definition of the do-operator holds just as well for specifiers for exogenous variables. The only difference is that the set PAR_Y is empty in those cases.

tions on the variables on the right-hand side. We now have an analogous invariance of specifiers in CBNs. As the first case of the definition shows, an intervention does not alter a specifier for a variable Y as long as the intervention is not on Y itself. This is still true, if the intervention is applied on parents of Y.

Notice that although the *do*-operator is defined as an operator on $\mathcal{P}$ in a Causal Bayes Net $(\mathcal{G}, \mathcal{P})$, it subsequently also induces a change in the graph $\mathcal{G}$. But the new graph $\mathcal{G}_{do(\vec{X}=\vec{x})}$ that results from the intervention $do(\vec{X} = \vec{x})$ cannot simply be obtained by constructing a graph that fulfills the Markov Condition relative to the new distribution $\mathcal{P}_{do(\vec{X}=\vec{x})}$. As pointed out before, the Markov Condition never tells us which arrows should not be included in a graph for a given probability distribution. It therefore cannot tell us that certain edges must be removed when certain dependencies are removed from the corresponding probability distribution. Since we already presupposed that we are dealing with causally adequate CBNs (or at least, with what a given agent considers to be causally adequate) and since the intervention $do(\vec{X} = \vec{x})$ is supposed to make the variables $\vec{X}$ causally independent from its parents, we obtain a causally adequate post-intervention graph $\mathcal{G}_{do(\vec{X}=\vec{x})}$ (or at least, one that a given agent considers to be causally adequate) from the pre-intervention graph $\mathcal{G}$ by deleting all directed edges that point to the variables in $\vec{X}$. So, ultimately, an intervention is an operator that modifies both the probability distribution $\mathcal{P}$ and the graph $\mathcal{G}$ of a CBN $(\mathcal{G}, \mathcal{P})$. I will write $(\mathcal{G}, \mathcal{P})_{do(\vec{X}=\vec{x})}$ to denote the CBN that results from applying the intervention $do(\vec{X} = \vec{x})$ on $(\mathcal{G}, \mathcal{P})$.

Having introduced interventions into CBNs, we can now also evaluate probabilistic intervention counterfactuals of the form: ‘the probability of the event ϕ would be p, if $\vec{X} = \vec{x}$ would be brought about by an intervention’, formally: $[\vec{X} \leftarrow \vec{x}]P(\phi) = p$. We say $(\mathcal{G}, \mathcal{P}) \models [\vec{X} \leftarrow \vec{x}]P(\phi) = p$ if and only if $(\mathcal{G}, \mathcal{P})_{do(\vec{X}=\vec{x})} \models P(\phi) = p$. We are also able to formulate more complex counterfactual formulas in our language, like: $[\vec{X} \leftarrow \vec{x}]P(\phi) > [\vec{Y} \leftarrow \vec{y}]P(\phi)$ (read as: the probability of ϕ after realizing $\vec{X} = \vec{x}$ by an intervention is higher than the probability of ϕ after realizing $\vec{Y} = \vec{y}$ by an intervention). We say $(\mathcal{G}, \mathcal{P}) \models [\vec{X} \leftarrow \vec{x}]P(\phi) > [\vec{Y} \leftarrow \vec{y}]P(\phi)$ if and only if there is a p, such that $(\mathcal{G}, \mathcal{P})_{do(\vec{X}=\vec{x})} \models P(\phi) = p$ and there is a q, such that $(\mathcal{G}, \mathcal{P})_{do(\vec{Y}=\vec{y})} \models P(\phi) = q$ and $p > q$.

2.3 Actual Causation in Causal Bayes Nets

2.3.1 Causal Bayes Nets and Token Events

As already pointed out, CBNs are a well entrenched tool for causal induction, that is, for deducing (probabilistic) causal relationships from a probability distribution that is understood to represent or summarize empirical or statistical data. For this task, the events in a CBN $(\mathcal{G}, \mathcal{P})$ (variables taking on certain values) are typically interpreted as event-types and not as singular or token events, while the probability distribution $\mathcal{P}$ is understood to be obtained from relative frequency data.[15] But if we aim to use CBNs to define the concept of actual (token) causation,

[15] To be clear, this does not mean that the probability distribution in a CBN must be understood as directly representing relative frequenties. What is important is only that the values of the distribution are guided by or based on the relative frequencies that are gained by empirical or statistical data. Hitchcock (2009a) similarly characterizes a probability distribution in a CBN by stressing: “*Pr* is a probability distribution over the variables in **V** that represents empirical probability, as estimated by frequency data” (Hitchcock, 2009a, p. 306). Neapolitan

we have to interpret the events in a CBN as token events, since it is token events that are the relata of actual causation. A simple frequency interpretation of $\mathcal{P}$ is thereby off the table, since token events are by definition non-repeatable. So, how shall we interpret the distribution $\mathcal{P}$ in a CBN $(\mathcal{G}, \mathcal{P})$ that is supposed to represent a token scenario? There are two main options.

Objective Single-Case Chances

The first option is this: The values of the probability distribution $\mathcal{P}$ in a CBN $(\mathcal{G}, \mathcal{P})$, that is supposed to represent a token scenario, represent objective single-case chances. We do not need to go into the details of what exactly objective single-case chances might be. It will suffice for our purposes to settle on a few features of single-case objective chances that are commonly accepted in the literature. (1) Single-case objective chances are ontic and non-epistemic, which means that the claim that an event E has a certain chance $\mathcal{P}(E)$ is either true or false in our actual world, inedependently of anyone's epistemic attitudes towards E. (2) Chances evolve over time. Especially, events that lie in the past only have trivial values of chance (1 or 0), since an event that lies in the past either did or did not happen. Any chance-function $\mathcal{P}$ is therefore relative to a given point in time t. (3) Especially according to so-called best-system accounts of objective chance,[16] objective single-case chances (of future events) are deeply connected to the probabilistic laws of our best scientific theories. As Lewis puts it: "the chances are what the probabilistic laws of the best system say they are" (Lewis, 1994, p. 480).

Although objective single-case chances are categorically different things than relative frequencies, a close relationship between relative frequencies and objective chances is typically assumed to obtain. This relationship can be summarized by the following formulation of the weak law of large numbers, as given by Briggs: "As the number of trials approaches infinity, the frequency of an outcome will converge on its chance with probability 1" (Briggs, 2010, p. 940). This means, if a token event E has the single-case chance p in a certain repeatable scenario $\mathfrak{S}$ and we repeat $\mathfrak{S}$ a large number of times, then we will observe that the relative frequency of E-type events (the number of occurrences of E in relation to the repetitions of $\mathfrak{S}$) will approach p. This is why objective chances are often understood to explain why certain repeatable events yield certain frequencies.[17] On the other hand, the intimate connection between objective chances and relative frequencies is especially useful, since the single-case objective chance of a token event E, if it does indeed exist, is obviously not directly observable. Data about the finite relative frequency of E-type events in adequate reference classes therefore often forms our only epistemic access to the single-case objective chance of a token event E.[18] This relationship between relative frequencies and single-case objective chances is also in accord with Lewis' (1994, p. 480) claim that objective single-case chances "are what the probabilistic laws of the best system say they are", since probabilistic laws are typically determined based on relative frequency data.

Still, the single-case objective chance of a token event E can significantly differ from the

and Jiang also point out: "The probability distribution in a Bayesian network is ordinarily based on the notion of a probability as a relative frequency" (Neapolitan and Jiang, 2017, p. 198).

[16] See, for example, (Lewis, 1994), (Hoefer, 2007), (Glynn, 2010). For an overview, see (Schwarz, 2016).

[17] See, for example, (Briggs, 2010), (Emery, 2015).

[18] As Briggs points out: "Frequencies are not an infallible guide to chances – an outcome's frequency can (with nonzero probability) diverge from its chance even over a large number of trials. Nonetheless, they are a good guide to chances" (Briggs, 2010, p. 938).

relative frequency of E-type events in a suitable reference class, depending on the time t, relative to which we measure the single-case objective chance of E. Since past events are commonly considered to have only trivial objective chances, the objective chance of a token event E that lies in the past has either the value 1 or the value 0, even if the relative frequency of bygone E-tokens in a suitable reference class has some non-trivial value between 0 and 1.

Credences

The second possible interpretation of the values of a probability distribution $\mathcal{P}$ in a CBN $(\mathcal{G}, \mathcal{P})$, that is supposed to represent a token scenario, are epistemic probabilities: In this case, the value of $\mathcal{P}(E)$ represents an agent's credence or degree of belief in the event E. $\mathcal{P}$ should therefore always be understood as being relative to a given agent α and to a certain time t, since epistemic states typically evolve over time. The credences of an agent α at time t are considered to be measurable and therefore epistemically accessible through the (hypothetical) betting behavior of α at time t.

Here again, although credences are categorically different things than relative frequencies, this does not mean that the degree of belief in a token event E and the relative frequency of E-type events are completely independent from each other. It is typically assumed that information about the relative frequency of E-type events in a suitable reference class should guide a rational agent's degree of belief in a token event E, at least as long as no trumping evidence overrides the informative value of the relative frequency of E-type events. Imagine someone tosses a coin about which we have the information that it has been tossed 10.000 times and it landed heads in 50% of the tosses. Directly after the toss the coin is hidden under a cup before we are able to see the outcome. At this point, believing that the coin landed heads (H) to a degree of 0.5 is the most rational thing to do. But as soon as we see the outcome of the toss with our own eyes, or a reliable witness tells us about the outcome, we have gained what I call *trumping evidence*, which directly bears on our credence in the token event H without bearing (in any significant way) on our credence about the relative frequency of H-type events.[19] So, in lack of any trumping evidence about whether a token event E of type $\mathbf{E}$ does or does not hold, a rational agent should follow what Hájek (2019) calls the *Principle of Direct Probability*:

The Principle of Direct Probability (PDP). *$\mathcal{P}(E|rel(\mathbf{E}) = p) = p$,*

where $\mathcal{P}$ is a credence function that measures the credences of a rational agent and $rel(\mathbf{E})$ is the relative frequency of the event-type $\mathbf{E}$ in a suitable reference class.

A very similar relationship between single-case objective chances and credences is commonly assumed. As long as there is no trumping evidence that overrides the informative value of the objective single-case chance of E, E's objective single-case chance should guide a rational agent's degree of belief in E.[20] David Lewis' Principal Principle is a famous version of this guideline

[19] The concept of trumping evidence is basically the same as Lewis' concept of *inadmissable evidence* in his formulation of the Principal Principle in (Lewis, 1986b). I only consider the term 'inadmissable' to be confusing, since it is surely admissible when it comes to forming rational degrees of beliefs. The crucial point is only that it invalidates the relationship between relative frequencies and credences as expressed in the Principle of Direct Probability and, as we will later see, the relationship between objective chances and credences as expressed in the Principal Principle.

[20] Since relative frequencies are typically seen as the empirical guide to objective single-case chances, this idea is rather straightforward, given the PDP.

and it can be summarized like this:

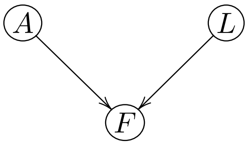
Principal Principle. $\mathcal{P}(E|ch(E) = p) = p$,

where $\mathcal{P}$ is a credence function that measures the credences of a rational agent and ch is an objective chance function.

This all indicates that even though objective chances of token events, credences about token events, and relative frequencies of repeatable event-types are ontologically clearly distinct interpretations of a probability distribution, the values of a given distribution can still systematically align under all three interpretations. But the conditions must be right for this. As soon as we consider past events, the objective single case chances detach from the relative frequencies of the corresponding event-types in suitable reference classes. And as soon as trumping evidence about token events are accessible for an agent, the agent's degrees of beliefs detach from the objective chances of the same events, as well as from the relative frequencies of the corresponding event-types in suitable reference classes. I will now argue that, if we want to use a CBN to represent the causal relationships in a token scenario, then we have to use a distribution, whose probability values (at least concerning the endogenous variables) align under all three interpretations.

Retaining Causal Information

Consider again the probabilistic forest fire scenario from section 2.2.2:



- $\mathcal{P}(A = 1) = 0.05$
- $\mathcal{P}(L = 1) = 0.1$

- $\mathcal{P}(F = 1|A = 1, L = 1) = 0.91$
- $\mathcal{P}(F = 1|A = 1, L = 0) = 0.7$
- $\mathcal{P}(F = 1|A = 0, L = 1) = 0.7$
- $\mathcal{P}(F = 1|A = 0, L = 0) = 0$

Figure 2.2: CBN for the probabilistic forest fire scenario.

As pointed out above, the specifications are assumed to conform to relative frequencies as obtained from statistical data and thereby encode all the quantitative information about the probabilistic causal relationships that hold in the scenario. Now, imagine that the CBN is used to represent a token scenario. The values of A represent, whether there has been an arsonist dropping a match in a specific forest on a specific day d. The values of L represent, whether there has been a lightning strike in the forest on d. And the values of F represent whether there has been a fire in the forest on d. Imagine also, that this specific scenario lies in the past, as is typically the case, when we try to find a causal explanation for an event, and imagine that $A = 1$, $L = 1$, and $F = 1$ actually happened. If we interpret $\mathcal{P}$ as an objective chance function relative to the present time t (or relative to any time t after d), we get:

- $\mathcal{P}_t(A = 1) = 1$
- $\mathcal{P}_t(L = 1) = 1$
- $\mathcal{P}_t(F = 1) = 1$

If we further assume, as is traditionally done, that the value of a conditional probability $\mathcal{P}(X|Y)$ is determined by the values of the unconditional probabilities $\mathcal{P}(Y)$ and $\mathcal{P}(X \wedge Y)$ through the ratio formula $\mathcal{P}(X|Y) = \dfrac{\mathcal{P}(X \wedge Y)}{\mathcal{P}(Y)}$, then we get the following specifiers:

- $\mathcal{P}_t(F = 1|A = 1, L = 1) = 1$
- $\mathcal{P}_t(F = 1|A = 1, L = 0) = undefined$
- $\mathcal{P}_t(F = 1|A = 0, L = 1) = undefined$
- $\mathcal{P}_t(F = 1|A = 0, L = 0) = undefined$

Clearly, those specifiers do not contain any information about the probabilistic causal relationships, as obtained from statistical data on relative frequencies. All information about the probabilistic causal relationships is completely lost.

But, as Hájek (2003) convincingly argues, there are good reasons for not following the classical view, according to which unconditional probabilities are primitive, while conditional probabilities are merely deduced from unconditional probabilities through the ratio formula. Instead, Hájek (2003), as well as several other authors,[21] favor a view according to which conditional probabilities are primitive and directly epistemically accessible, while the ratio formula is not seen as a definition of conditional probabilities in terms of unconditional probabilities, but merely as a "constraint on conditional probability: when $\mathcal{P}(A \wedge B)$ and $\mathcal{P}(B)$ are both sharply defined, and $\mathcal{P}(B)$ is non-zero, the probability of A given B is constrained to be their ratio" (Hájek, 2003, p. 276. *Notation adjusted*). But as soon as the mentioned conditions are not met, "there is no constraining to be done by the ratio" (Hájek, 2003, p. 276).

Adopting this view makes a clear difference to the values of the specifiers $\mathcal{P}_t(F = 1|A = 1, L = 0)$, $\mathcal{P}_t(F = 1|A = 0, L = 1)$, and $\mathcal{P}_t(F = 1|A = 0, L = 0)$. Since the conditions that validate the ratio formula are not fulfilled in these cases, the ratio formula does not constrain what values these conditional probabilities can take on. The values that align with our data on relative frequencies, namely $\mathcal{P}_t(F = 1|A = 1, L = 0) = 0.7$, $\mathcal{P}_t(F = 1|A = 0, L = 1) = 0.7$, $\mathcal{P}_t(F = 1|A = 0, L = 0) = 0$, are therefore legitimate. But a major problem remains. For $\mathcal{P}_t(F = 1|A = 1, L = 1)$, the conditions that validate the ratio formula are fulfilled, so the ratio formula does constrain the value of $\mathcal{P}_t(F = 1|A = 1, L = 1)$. And since $\mathcal{P}(A = 1 \wedge L = 1 \wedge F = 1) = \mathcal{P}(F = 1) = 1$, we still get $\mathcal{P}_t(F = 1|A = 1, L = 1) = 1$. This still conflicts with the known strength of the causal relationship between $A = 1$ and $F = 1$, as well as the one between $L = 1$ and $F = 1$. According to our empirically gained data, it was absolutely compatible with the known causal relationships, that $F = 1$ would not have happened, even though $A = 1$ and $L = 1$

[21] See for example Climenhaga (2020), who argues for taking conditional probabilities in a CBN as basic and unconditional probabilities as derivative. Also, Popper (1959) directly axiomatizes conditional probabilities as basic probabilities. See, for example, (Leitgeb, 2012) for a concise summary of Popper-functions.

did happen. This does not hold for the distribution $\mathcal{P}_t$. Put differently, $\mathcal{P}_t$ does not provide us with the correct value of the probability that $F = 1$ would have had, if we would have intervened to realize $A = 1$ and $L = 1$.

The only way to solve this issue, is to use an objective chance function relative to a time t_0 that lies before the time of the considered causal scenario, which means t_0 should not be later than any event that is represented by a root-variable of the CBN in question. By this, we ensure that the specifiers for the endogenous variables still align with our probabilistic laws and our relative frequency data. Imagine that in the token forest fire scenario $A = 1$ and $L = 1$ happened at exactly the same time t_0, while $F = 1$ happened a bit later. We can then take an objective chance function $\mathcal{P}_{t_0}$ relative to t_0, which gives us $\mathcal{P}_{t_0}(A = 1) = 1$ and $\mathcal{P}_{t_0}(L = 1) = 1$. But since $F = 1$ did not yet happen at t_0, our relative frequency data guides the assignment of the following objective chance values:

- $\mathcal{P}_{t_0}(F = 1|A = 1, L = 1) = 0.91$
- $\mathcal{P}_{t_0}(F = 1|A = 1, L = 0) = 0.7$
- $\mathcal{P}_{t_0}(F = 1|A = 0, L = 1) = 0.7$
- $\mathcal{P}_{t_0}(F = 1|A = 0, L = 0) = 0$

Unsurprisingly, these specifiers adequately encode the known probabilistic causal relationships of the scenario.

Let us now imagine that $\mathcal{P}$ is interpreted as a credence function, while the causal scenario again lies in the past. As long as we do not have any trumping evidence about which values the variables actually took on, we only have our relative frequency data to guide our credence function, such that:

- $\mathcal{P}(A = 1) = 0.05$
- $\mathcal{P}(L = 1) = 0.1$
- $\mathcal{P}(F = 1|A = 1, L = 1) = 0.91$
- $\mathcal{P}(F = 1|A = 1, L = 0) = 0.7$
- $\mathcal{P}(F = 1|A = 0, L = 1) = 0.7$
- $\mathcal{P}(F = 1|A = 0, L = 0) = 0$

But now imagine that we get hold of some trumping evidence about the actual value of F. We might, for example, come across an old newspaper article which states that $F = 1$ happened. Assuming that this evidence is trustworthy, this gain of knowledge clearly influences our credences about the scenario. This influence is typically represented by updating the credence function through conditionalization on $F = 1$. With $\mathcal{P}' = \mathcal{P}(\cdot|F = 1)$, we get: $\mathcal{P}'(F = 1|A = 1, L = 1) = 1$, $\mathcal{P}'(F = 1|A = 1, L = 0) = 1$, $\mathcal{P}'(F = 1|A = 0, L = 1) = 1$.[22] So

[22] This follows from the ratio formula and the fact that $\mathcal{P}'(F = 1) = 1$. Notice that the ratio formula is a valid constraint on the mentioned conditional probabilities, since the conditions for its validity are fulfilled in the given cases.

here again, the information about the probabilistic causal relationships, that hold in the given scenario, is completely lost. To avoid this loss of information, we have to keep out any trumping evidence about the actual values of the endogenous variables. This way, we ensure that the conditional credences about the values of the endogenous variables do align with our relative frequency data, which contains the information about the probabilistic causal relationships, since the relative frequency data is then the only evidence left, that can guide our credences.

Let us take stock. We have seen that if we want to use a CBN $(\mathcal{G}, \mathcal{P})$ to represent a token scenario, the probability distribution $\mathcal{P}$ cannot simply be understood as representing relative frequencies, as is typically done in CBNs that represent event types. Instead, we face two options. $\mathcal{P}$ can either be interpreted as an objective chance function or as a credence function. But these interpretations come with some problems. An objective chance function assigns only trivial probabilities to past events and thereby forfeits its ability to adequately encode probabilistic causal relationships between bygone token events. The same happens to a credence function, as soon as trumping evidence about the actual values of endogenous variables comes into play. To ensure that an objective chance function still accurately represents the probabilistic causal relationships between past events, we have to use an objective chance function relative to a time t that is not later than any event represented by a root-variable of the CBN. And to ensure that a credence function accurately encodes the probabilistic causal relationships between the events in a CBN, we have to keep out any trumping evidence about the actual values of the endogenous variables in the CBN. If we follow these rules, then the specifiers of $\mathcal{P}$ in the CBN $(\mathcal{G}, \mathcal{P})$ still encodes all the relevant information about the probabilistic causal relationships that are supposed to hold in the presented scenario, no matter, whether we interpret $\mathcal{P}$ as an objective chance function or as a credence function.

Actual Events for Actual Causation

There is another lesson we can learn from the last chapter: A CBN $(\mathcal{G}, \mathcal{P})$, which represents a past token scenario, does in general not contain both: adequate information about the probabilistic causal relationships between the events in the scenario and information about which of those events actually happened. Whenever we want to retain information about the probabilistic causal relationships between the events in the scenario, we either have to use an objective chance function $\mathcal{P}$ relative to a time t before the considered scenario took place or a credence function $\mathcal{P}$ that does not allow any trumping evidence about the values of endogenous variables. In both cases, the resulting probabilities over the values of the endogenous variables do, in general, not give an accurate description of which values the variables actually took on.

This amounts to a problem when we want to define actual causation in a CBN. A crucial condition in the HP-definition of actual causation in deterministic causal scenarios is that the actual cause $\vec{X} = \vec{x}$ and the effect ϕ must actually have happened. This condition is just as intuitive in a situation with probabilistic causal relationships. So, if we want to hold on to the idea that the relation of actual causation can only obtain between events that actually happen, we need to refer to information that a CBN can, in general, not provide us with.

A possible solution to this problem is to enrich the classical CBN-framework with a second probability distribution $\mathcal{C}_{pr}$ that does not represent the causal relationships between the events but only the objective chances of the events in the model relative to the present time. Instead of defining actual causation relative to a classical CBN that consists of a tuple $(\mathcal{G}, \mathcal{P})$, we simply define actual causation relative to the triple $(\mathcal{G}, \mathcal{P}, \mathcal{C}_{pr})$ where $\mathcal{P}$ encodes the information about the probabilistic causal relationships between the events in the considered scenario and $\mathcal{C}_{pr}$ represents which events actually did or did not happen. We can then formulate the first condition for $\vec{X} = \vec{x}$ to be an actual cause of ϕ in the following way:

- $\mathcal{C}_{pr}(\vec{X} = \vec{x}) = 1$ and $\mathcal{C}_{pr}(\phi) = 1$.

But what about the other conditions of actual causation? How can we transfer the remaining conditions of actual causation into the framework of CBNs?

2.3.2 Probability Change Instead of De Facto Dependence

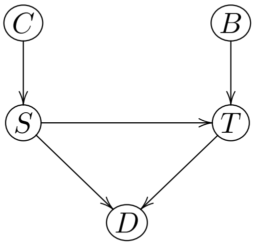
The original HP-definition of actual causation requires that an effect must be de facto dependent on its actual cause, in the sense specified by the condition AC2. But this kind of de facto dependence cannot be guaranteed by probabilistic causal relationships. Consider again the probabilistic forest fire scenario, in which $A = 1$, $L = 1$, and $F = 1$ actually happened. There is indeed a contingency, in which we can yield $F = 0$ by setting A to 0, namely the contingency in which L is set to 0 as well. But we have a problem, when it comes to condition AC2(b). It must hold in the same contingency that F would take on value 1, if A is set to 1 by an intervention. But the probabilistic nature of the causal connection between the two variables does not guarantee this result, even if $A = 1$ did causally yield $F = 1$ in the actual situation. The probabilistic causal relationship between $A = 1$ and $F = 1$ only guarantees that the probability of $F = 1$ will be 0.7, if A is set to 1 by an intervention in that contingency. This suggests that the heart of the criterion of actual causation in deterministic causal scenarios needs a fundamental reformulation to make it applicable to scenarios with probabilistic causal relationships.

The natural analogue to a counterfactual dependence of the occurrence of the effect on the occurrence of the cause in the deterministic case, is a counterfactual dependence of the effect's probability on the occurrence of the cause. It is, of course, a well known fact, that probabilistic dependence is neither necessary nor sufficient for actual causation, at least as long as we try to capture the idea of probabilistic dependence with conditional probabilities. Conditioning on the effect increases the probability of the cause, just like conditioning on the cause increases the probability of the effect. But the concept of an intervention provides us with a great tool to solve this problem of causal asymmetry: Realizing the cause by an intervention increases the probability of the effect, but realizing the effect by an intervention does not increase the probability of the cause. This provides us with a first straightforward proposal to replace the condition of counterfactual dependence that we have used for defining actual causation in deterministic causal scenarios: For $\vec{X} = \vec{x}$ to be an actual cause of ϕ, $\vec{X} = \vec{x}$ must yield an increase of ϕ's probability in the following way:

Probability Raising. $\mathcal{P}_{(\vec{X}=\vec{x})}(\phi) > \mathcal{P}_{(\vec{X}=\vec{x}')}(\phi)$ *for some alternative values* $\vec{x}' \neq \vec{x}$ *of* $\vec{X}$.

But just like with simple counterfactual dependence in deterministic causal scenarios, *Probability Raising* itself is not yet good enough to define actual causation in probabilistic causal scenarios.

While simple counterfactual dependence turns out to be unnecessary for actual causation in deterministic causal scenarios due to cases of preemption, *Probability Raising* is neither necessary nor sufficient for actual causation. To see this, consider an example that has been put forward by Fenton-Glynn (2017). Despite being a bit more sinister, the example is structurally quite similar to the bottle breaking scenario from chapter 1. The story goes like this: Two mafia bosses, Don Corleone and Don Barzini, want to kill Police Chief McCluskey. Both order their respective hitmen, Sonny and Turk, on the very same day. Both hitmen are quite, but not totally, reliable. So, when Corleone orders Sonny to shoot McCluskey ($C = 1$), there is a chance of 0.9 that Sonny obeys and shoots McCluskey ($S = 1$). Similarly, when Barzini orders Turk to shoot McCluskey ($B = 1$), there is a chance of 0.9 that Turk obeys and shoots the Police Chief ($T = 1$). The mafia bosses also had the chance to gather a big amount of data to calculate the success rate of their hitmen. Sonny has a shaky hand and only kills half of the time when he shoots. So with no other potential causes of McCluskey's death ($D = 1$) present, we have: $\mathcal{P}(D = 1 \mid S = 1 \wedge T = 0) = 0.5$. Turk, on the other hand, is much more successful in his trade. He has a success rate of 0.9. So, with no other causes of McCluskey's death present, we have: $\mathcal{P}(D = 1 \mid T = 1 \wedge S = 0) = 0.9$. Now, here comes a twist: Even though it is his job, Turk does not really like to kill people. He therefore has the habit to wait for his competitor Sonny to do the job. Turk only shoots, if Sonny does not shoot, which means: $\mathcal{P}(T = 1 \mid S = 1, B = 1) = \mathcal{P}(T = 1 \mid S = 1, B = 0) = 0$. We can represent this situation by a CBN $(\mathcal{G}, \mathcal{P})$ as presented in figure 2.3.



- $\mathcal{P}(S = 1 \mid C = 1) = 0.9$
- $\mathcal{P}(S = 1 \mid C = 0) = 0$
- $\mathcal{P}(T = 1|S = 1, B = 1) = 0$
- $\mathcal{P}(T = 1|S = 0, B = 1) = 0.9$
- $\mathcal{P}(T = 1|S = 1, B = 0) = 0$
- $\mathcal{P}(T = 1|S = 0, B = 0) = 0$
- $\mathcal{P}(D = 1|S = 1, T = 1) = 0.95$
- $\mathcal{P}(D = 1|S = 0, T = 1) = 0.9$
- $\mathcal{P}(D = 1|S = 1, T = 0) = 0.5$
- $\mathcal{P}(D = 1|S = 0, T = 0) = 0$

Figure 2.3: CBN for the mafia scenario.

Now, here is what actually happened: Both, Corleone and Barzini gave the order to kill McCluskey. So, we have: $C = 1$ and $B = 1$. Sonny obeyed and shot McCluskey ($S = 1$), who died ($D = 1$). Turk did not shoot ($T = 0$). Intuitively, $C = 1$ is clearly an actual cause of $D = 1$, while $B = 1$ is not. But $C = 1$ does not satisfy *Probability Raising* in regard to $D = 1$, because $\mathcal{P}_{do(C=1)}(D = 1) > \mathcal{P}_{do(C=0)}(D = 1)$ does not hold.

Proof. First, $\mathcal{P}_{do(C=1)}(D=1)=0.531$, because:
We have $\mathcal{P}_{do(C=1)}(C=1)=1$, $\mathcal{P}_{do(C=1)}(S=1)=0.9$. We further have: $\mathcal{P}_{do(C=1)}(T=1 \mid S=1)=0$, which gives us $\mathcal{P}_{do(C=1)}(T=0 \mid S=1)=1$ and $\mathcal{P}_{do(C=1)}(T=1|S=0)=0.9\times P(B=1)=0.9$. We therefore get: $\mathcal{P}_{do(C=1)}(D=1)=0.95\times\mathcal{P}_{do(C=1)}(T=1|S=1)\times\mathcal{P}_{do(C=1)}(S=1)+0.9\times\mathcal{P}_{do(C=1)}(T=1|S=0)\times\mathcal{P}_{do(C=1)}(S=0)+0.5\times\mathcal{P}_{do(C=1)}(T=0|S=1)\times\mathcal{P}_{do(C=1)}(S=1)=0.9\times0.9\times0.1+0.5\times0.9=0.531$.

But, we have $\mathcal{P}_{do(C=0)}(D=1)=0.81$, because:
We have $\mathcal{P}_{do(C=0)}(C=1)=0$ and therefore $\mathcal{P}_{do(C=0)}(S=1)=0$. So, $\mathcal{P}_{do(C=0)}(D=1)=\mathcal{P}_{do(C=0)}(D=1|S=0,T=1)=0.9\times\mathcal{P}_{do(C=0)}(T=1|S=0)\times\mathcal{P}_{do(C=0)}(S=0)=0.9\times0.9\times1=0.81$. □

Probability Raising can therefore not be a necessary condition for actual causation. But it can also not be a sufficient condition for actual causation, because $B=1$, which is intuitively clearly no actual cause of $D=1$, satisfies *Probability Raising* relative to $D=1$.

Proof. First, $\mathcal{P}_{do(B=1)}(D=1)=0.531$, because:
We have $\mathcal{P}_{do(B=1)}(B=1)=1$, we have $\mathcal{P}_{do(B=1)}(C=1)=1$, $\mathcal{P}_{do(B=1)}(S=1)=0.9\times1=0.9$, and $\mathcal{P}_{do(B=1)}(T=1|S=1)=0$, $\mathcal{P}_{do(B=1)}(T=0|S=1)=1$ and $\mathcal{P}_{do(B=1)}(T=1|S=0)=0.9$. So we have: $\mathcal{P}_{do(B=1)}(D=1)=0.95\times\mathcal{P}_{do(B=1)}(T=1|S=1)\times\mathcal{P}_{do(B=1)}(S=1)+0.9\times\mathcal{P}_{do(B=1)}(T=1|S=0)\times\mathcal{P}_{do(B=1)}(S=0)+0.5\times\mathcal{P}_{do(B=1)}(T=0|S=1)\times\mathcal{P}_{do(B=1)}(S=1)=0.9\times0.9\times0.1+0.5\times0.9=0.531$.

And we have $\mathcal{P}_{do(B=0)}(D=1)=0.45$, because:
We have $\mathcal{P}_{do(B=0)}(B=1)=0$ and therefore $\mathcal{P}_{do(B=0)}(T=1)=0$. So, $\mathcal{P}_{do(B=0)}(D=1)=\mathcal{P}_{do(B=0)}(D=1|S=1,T=0)=0.5\times\mathcal{P}_{do(B=0)}(T=0|S=1)\times\mathcal{P}_{do(B=0)}(S=1)=0.5\times1\times0.9=0.45$.

Therefore: $\mathcal{P}_{do(B=1)}(D=1)>\mathcal{P}_{do(B=0)}(D=1)$. □

2.3.3 Fenton-Glynn's PC-Definition

The HP-definition of actual causation has the condition of counterfactual dependence at its core. But Halpern and Pearl had to acknowledge the fact that counterfactual dependence itself is not necessary for actual causation. This is why condition AC2 of the HP-definition does not demand a simple counterfactual dependence of the effect on the cause, but a de facto dependence, which means: The effect does not have to counterfactually dependent on the cause in the actual causal setting, but only in a certain contingency. Fenton-Glynn proposes that pretty much the same trick can be applied to the condition of *Probability Raising* to obtain an appropriate definition of actual causation in probabilistic causal scenarios. The basic idea is this: The cause does not have to raise the probability of its effect in the actual causal setting, but only in a certain contingency. In analogy to the concept of de facto dependence, we can call this kind of probability raising 'de facto probability raising'. Based on this idea, Fenton-Glynn (2017, cf. p. 1098) gives the following explication of actual causation in the framework of CBNs:[23]

[23] Fenton-Glynn's notation is slightly but insignificantly different than mine. The most obvious difference is my use of $\mathcal{C}_{pr}$. Although Fenton-Glynn makes a reference to events in the actual world in condition PC1, he does not use a distribution $\mathcal{C}_{pr}$ as a formal tool to represent which events actually happened.

Actual Cause (PC) (Fenton-Glynn, 2017). *$\vec{X} = \vec{x}$ is an actual cause of ϕ in $(\mathcal{G}, \mathcal{P}, \mathcal{C}_{pr})$ if and only if the following conditions hold:*

PC1 *$\mathcal{C}_{pr}(\vec{X} = \vec{x}) = 1$ and $\mathcal{C}_{pr}(\phi) = 1$*

PC2 *There is a partition $(\vec{Z}, \vec{W})$ of the endogenous variables $\mathcal{V}$ in $\mathcal{G}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w})$ such that if $\mathcal{C}_{pr}(Z = z^*) = 1$ for all $Z \in \vec{Z}$, then:*

- *$(\mathcal{G}, \mathcal{P}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]P(\phi) > [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]P(\phi)$ for all $\vec{Z}' \subseteq \vec{Z}$.*

PC3 *$\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies conditions PC1 and PC2.*

Fenton-Glynn's definition is clearly reminiscent of Halpern and Pearl's definition of actual causation in deterministic causal models. The crucial difference is of course, that Fenton-Glynn uses the condition of *Probability Raising* in place of counterfactual dependence in condition PC2. Notice further that the two conditions of AC2(a) and AC2(b) in the HP-definition merge into one condition in PC2. This is because the comparison between the effects of $\vec{X} = \vec{x}$ and $\vec{X} = \vec{x}'$ is now captured in a single inequality.[24] Another difference is that the condition formulated in PC2 does not need to hold for any subset $\vec{W}'$ of $\vec{W}$ as in the condition in AC2 of the HP-definition. Fenton-Glynn points out that this demand would lead to some problems in the probabilistic case, while abandoning it seems to be unproblematic.[25]

Now, the question is, does the same idea that worked for the condition of counterfactual dependence in the deterministic case, also work for the condition of *Probability Raising* in probabilistic scenarios? The mafia scenario from above provides a good initial test. First of all, in accordance with intuition, $C = 1$ is indeed an actual cause of $D = 1$ in the mafia scenario according to PC.

Proof. Condition PC1 is obviously fulfilled, because in our example we presupposed that $C = 1$ and $D = 1$ are the case. PC3 is trivially fulfilled, since $C = 1$ is a singleton. So, it all depends on PC2. To show that PC2 is fulfilled, we simply choose the contingency in which Barzini gives his order to kill, but Turk does not obey, so $\vec{W} = \vec{w}$ consists of $B = 1$ and $T = 0$, while $\vec{Z} = \vec{z}^*$ consists of $C = 1$, $S = 1$, and $D = 1$. So, we have to check, whether $[C \leftarrow 1, B \leftarrow 1, T \leftarrow 0, \vec{Z}' \leftarrow \vec{z}^*]P(D = 1) > [C \leftarrow 1, B \leftarrow 1, T \leftarrow 0]P(D = 1)$ is true in $(\mathcal{G}, \mathcal{P})$ for all $\vec{Z}' \subseteq \vec{Z}$. For $\vec{Z}' = \varnothing$, we have $\mathcal{P}_{(C=1,B=1,T=0)}(D = 1) = 0.45$ (since T is set to 0, $\mathcal{P}_{(C=1,B=1,T=0)}(D = 1) = 0.5 \times \mathcal{P}_{(C=1,B=1,T=0)}(T = 0|S = 1) \times \mathcal{P}_{(C=1,B=1,T=0)}(S = 1) = 0.5 \times 1 \times 0.9$). For $\vec{Z}' = \vec{z}^*$ being $C = 1$ the result is obviously the same. For $\vec{Z}' = \vec{z}^*$ being $C = 1 \wedge S = 1$ and $\vec{Z}' = \vec{z}^*$ being $S = 1$ the result is even higher, since we have $\mathcal{P}_{(C=1,B=1,T=0,S=1)}(D = 1) = 0.5$ (again, since T is set to 0 and S is set to 1, $\mathcal{P}_{(C=1,B=1,T=0)}(D = 1) = 0.5 \times \mathcal{P}_{(C=1,B=1,T=0)}(T = 0|S = 1) \times \mathcal{P}_{(C=1,B=1,T=0)}(S = 1) = 0.5 \times 1 \times 1$.) Finally, for $\vec{Z}' = \vec{z}^*$ being $C = 1 \wedge S = 1 \wedge D = 1$ or any other conjunction that includes $D = 1$, we obviously have the result that $\mathcal{P}_{do(C=1,B=1,T=0,S=1,D=1)}(D = 1) = 1$. Now, for PC2 to be

[24] For reasons of analogy to the deterministic defintion of actual causation, Fenton-Glynn separated condition PC2 into two subparts. But this is in no way necessary.

[25] Actually, in an earlier version of the HP-definition, Halpern and Pearl did not include this condition in the deterministic case either. Fenton-Glynn is convinced that this version is actually preferable. I will not go into the details of his argument, since it will not affect the following discussion.

satisfied, all these results (0.45, 0.5 and 1) must be bigger than $\mathcal{P}_{(C=0,B=1,T=0)}(D=1)$. This is indeed the case, since $\mathcal{P}_{(C=0,B=1,T=0)}(D=1)=0$: We have $\mathcal{P}_{do(C=0,B=1,T=0)}(S=1)=0$ and $\mathcal{P}_{do(C=0,B=1,T=0)}(T=1)=0$. So, we have: $\mathcal{P}_{do(C=0,B=1,T=0)}(D=1)=0.95\times 0\times 0+0.9\times 0\times 1+0.5\times 1\times 0=0$. Therefore, $C=1$ is an actual cause of $D=1$ according to PC. □

Furthermore, $B=1$ does not satisfy PC in regard to $D=1$, which means that PC also correctly recognizes that $B=1$ is no actual cause of $D=1$.

Proof. T must be in $\vec{Z}$, since otherwise T would be in $\vec{W}$, where it would screen off any causal influence from $B=1$ to $D=1$, no matter if we would set T to 1 or to 0. But if T is in $\vec{Z}$, then for satisfying PC2, it must, for example, be the case that $\mathcal{P}_{do(B=1,\vec{W}=\vec{w},T=0)}(D=1) > \mathcal{P}_{do(B=0,\vec{W}=\vec{w})}(D=1)$. But this does not hold, since we have: $\mathcal{P}_{do(B=1,\vec{W}=\vec{w},T=0)}(D=1) = \mathcal{P}_{do(B=0,\vec{W}=\vec{w})}(D=1)$. With T being set to 0, setting B to 1 does not make any difference to the probability of $D=1$ in comparison to the situation, in which the value of T is not fixed to any value and B is set to 0. □

2.3.4 Actual Causation as De Facto Probability Raising

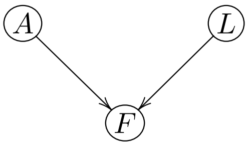
Fenton-Glynn's definition of actual causation looks promising. It transfers the central idea of the HP-definition into probabilistic causal contexts and thereby enables an explication of actual causation in a framework, which has proved very useful in representing probabilistic causal relationships: Causal Bayes Nets.

In the introduction to this chapter, I have already introduced two different views on what actual causation amounts to in the context of probabilistic causal relationships. According to the first view, which I called the de facto probability raising account, actual causation in the context of probabilistic causal relationships is nothing else than the cause C producing a certain increase of the probability of the effect E. According to the second view, which I dubbed the de facto dependence account, a cause candidate either does causally produce and therefore yield the effect E in question or it does not. Probabilities only come into play on the type-level, when only a certain percentage of a set of C-tokens produces E-tokens, or on an epistemic level, when there is an uncertainty about whether a given C-token ultimately does or did produce E or not. According to the first view, it is the production of probabilities of token events that ground probabilistic causal relationships on the type-level. According to the second view, it is the production of token-events itself that ground probabilistic causal relationships on the type-level.

Fenton-Glynn's PC-definition of actual causation is clearly a de facto probability raising account. The condition of de facto probability raising is the core of his definition PC.

To see that the differentiation between the two types of accounts of actual causation is not just some ontological quibble, but has actual consequences for the recognition and classification of cases of actual causation, consider again our probabilistic forest fire scenario as presented in figure 2.4.

We again consider an actual situation, in which $A=1$, $L=1$ and $F=1$ happened. Now, what does PC say about the relation between $A=1$ and $F=1$? First of all, $A=1$ satisfies condition PC1, since $\mathcal{C}_{pr}(A=1)=1$ and $\mathcal{C}_{pr}(F=1)=1$. $A=1$ also satisfies PC3, since it is a singleton. To show, that $A=1$ satisfies PC2, we choose $\vec{W}=\{L\}$ and $\vec{w}=0$. Since only A and F are



- $\mathcal{P}(A = 1) = 0.05$
- $\mathcal{P}(L = 1) = 0.1$

- $\mathcal{P}(F = 1|A = 1, L = 1) = 0.91$
- $\mathcal{P}(F = 1|A = 1, L = 0) = 0.7$
- $\mathcal{P}(F = 1|A = 0, L = 1) = 0.7$
- $\mathcal{P}(F = 1|A = 0, L = 0) = 0$

Figure 2.4: CBN for the probabilistic forest fire scenario.

in $\vec{Z}$, the only nontrivial inequality that we have to check, is $[A \leftarrow 1, L \leftarrow 0]P(F = 1) > [A \leftarrow 0, L \leftarrow 0]P(F = 1)$. Since we have $\mathcal{P}_{do(A=1,L=0)}(F = 1) = 0.7$ and $\mathcal{P}_{do(A=0,L=0)}(F = 1) = 0$, we have $(\mathcal{G}, \mathcal{P}) \models [A \leftarrow 1, L \leftarrow 0]P(F = 1) > [A \leftarrow 0, L \leftarrow 0]P(F = 1)$. $A = 1$ is therefore an actual cause of $F = 1$ in $(\mathcal{G}, \mathcal{P})$, according to PC. Analogously it can be shown that PC also recognizes $L = 1$ as an actual cause of $F = 1$ in $(\mathcal{G}, \mathcal{P})$. So, PC classifies $A = 1$ and $L = 1$ as actual causes of $F = 1$ in the probabilistic forest fire scenario, just like it classifies $C = 1$ as an actual cause of $D = 1$ in the mafia scenario. For a de facto probability raising account of actual causation, there is no significant difference between both examples, when it comes to actual causation. In both cases, the cause-candidate produces an increase of the probability of the effect and the effect actually happens. This is all that counts for actual causation according to this view.

But as soon as we adopt a de facto dependence interpretation of actual causation, we notice a crucial difference between both examples. Given the knowledge about the causal relationships and the actual values of the variables in the mafia scenario, we know for sure that the event $C = 1$ kicked of a causal process that ultimately causally produced the death of McCluskey. The type-level causal relationship between $C = 1$ and $D = 1$ is non-deterministic. We therefore know that there was a possibility that $C = 1$ happened without succeeding to produce $D = 1$. But in the given scenario, this possibility did not actualize, because otherwise McCluskey could not have died. The only alternative causal process that could have produced $D = 1$ goes via $T = 1$. So, knowing that $T = 1$ did not happen while $D = 1$ did happen, means that $C = 1$ successfully produced $D = 1$ in the actual situation. A de facto dependence interpretation of actual causation would therefore also classify $C = 1$ as an actual cause of $D = 1$ in the mafia scenario

But now consider the forest fire scenario. The type-level causal relationship between $A = 1$ and $F = 1$ leaves a 30% chance that the token event $A = 1$ does not succeed in producing $F = 1$. The same holds for the token event $L = 1$. We know that $F = 1$ actually happened, which means that either $A = 1$, $L = 1$ or both must have causally produced $F = 1$ in the actual situation. But our knowledge about the situation does not allow us to conclusively say which one it was. So, with the knowledge that is given about the situation, there remains an uncertainty about whether $A = 1$ (or $L = 1$) succeeded in causally producing $F = 1$. But according to the de facto dependence account it is just this success in the causal production of $F = 1$ that would make $A = 1$ (and likewise $L = 1$) an actual cause of $F = 1$. Unlike PC a de facto dependence account of actual causation would therefore not conclusively classify $A = 1$ (or $L = 1$) as an actual cause of $F = 1$ in the probabilistic forest fire scenario.

So far, we have only differentiated a de facto dependence account of actual causation from Fenton-Glynn's approach. But the question remains, how exactly a de facto dependence account of actual causation can be spelled out in the context of probabilistic causal relationships. It is now time to answer that question.

2.4 The SEM-Approach to Probabilistic Causal Models

2.4.1 Probabilistic Structural Equation Models

To represent probabilistic causal relationships we have put aside structural equation models (SEMs) with their deterministic structural equations and turned to CBNs at the beginning of this chapter. With that move, we may have thrown out the baby with the bathwater. Despite the inherently deterministic character of structural equations, SEMs can actually be amended to represent probabilistic causal relationships, namely by embadding error-terms into the structural equations. Halpern (2016) and Pearl (2000) both put forward this approach as an alternative to a CBN-modeling of scenarios with probabilistic causal relationships. To see how this approach works, consider again our forest fire examples: In the deterministic case, in which $A = 1$ and $L = 1$ both deterministically yield $F = 1$ independently from each other, we used the structural equation $F := A \vee L$ to represent these causal relationships. Consider now the probabilistic scenario, in which $A = 1$ yields $F = 1$ with a probability of 0.7 when no other causal influence on variable F is present and $L = 1$ also yields $F = 1$ with a probability of 0.7 when no other causal influence on variable F is present. To model these probabilistic causal relationships, we amend the deterministic structural equation model with two additional exogenous variables, $U_{A,F}$ and $U_{L,F}$, that can both take on two different values. $U_{A,F} = 1$ represents that, given that $A = 1$ is the case, $A = 1$ successfully causally produces $F = 1$. $U_{A,F} = 0$ represents that, given that $A = 1$ is the case, the causal production of $F = 1$ by $A = 1$ does not succeed. Similarly, $U_{L,F} = 1$ represents that, given that $L = 1$ is the case, $L = 1$ successfully causally produces $F = 1$. $U_{L,F} = 0$ represents that, given that $L = 1$ is the case, the causal production of $F = 1$ by $L = 1$ does not succeed. The probabilistic causal relationship between A and F, as well as the one between L and F, can now be represented by two components: The structural equation $F := (A \wedge U_{A,F}) \vee (L \wedge U_{L,F})$ and a probability distribution $\mathcal{P}$ over the values of the error-terms with $\mathcal{P}(U_{A,F} = 1) = 0.7$ and $\mathcal{P}(U_{L,F} = 1) = 0.7$. Notice that the structural equation is still deterministic. It tells us that $F = 1$ happens either if $A = 1$ and $U_{A,F} = 1$ are the case or $L = 1$ and $U_{L,F} = 1$ are the case. The probabilistic character of the causal relationships is expressed in form of the probability distribution over the error-terms.

Just as with CBNs, there are different possibilites to interpret the probabilities in an SEM. If we use an SEM to represent the causal relationships between event-types, then we can interpret the probabilities given by $\mathcal{P}$ as relative frequencies. $\mathcal{P}(U_{A,F} = 1) = 0.7$, for example, would mean that in a given data set, 70% of all instances, in which $A = 1$ happens, $A = 1$ produces

$F = 1$. If we, on the other hand, aim to represent a token scenario with an SEM, then we might either interpret the values of $\mathcal{P}$ as objective single-case chances or as agent-relative credences. Interpreted as an objective single-case chance, $\mathcal{P}(U_{A,F} = 1) = 0.7$, for example, means that, given that the token event $A = 1$ happens, there would be an objective single case chance of 70% that the token event $A = 1$ produces the token event $F = 1$. Interpreted as a degree of belief, $\mathcal{P}(U_{A,F} = 1) = 0.7$ means that, assuming that the token event $A = 1$ happens, a given agent would belief to a degree of 0.7, that the token event $A = 1$ produces the token event $F = 1$.[26]

Remember that a deterministic SEM $\mathcal{M}$ consists of a set $\mathcal{V}$ of endogenous variables, a set $\mathcal{U}$ of exogenous variables, a range $\mathcal{R}$ that assigns every variable in $\mathcal{V} \cup \mathcal{U}$ a set of values and a set $\mathcal{F}$ of structural equations, one for each endogenous variable $X \in \mathcal{V}$. The same still holds for an SEM that is used to represent probabilistic causal relationships. The only difference is that the set $\mathcal{U}$ of exogenous variables can be differentiated into two subgroups, a set $\mathcal{U}_e$ of error-terms and a set $\mathcal{U}_b$ of what I will call background variables. Since both are exogenous variables, they both only appear on the right hand-side of the structural equations in $\mathcal{M}$. But a background variable never appears together with an endogenous variable on the right hand side of a structural equation in $\mathcal{M}$. Error-terms, on the other hand, only appear with endogenous variables on the right hand side of structural equations in $\mathcal{M}$. Just like in SEMs, that only represent deterministic causal relationships, we have contexts for causal models, where a context $\vec{u}$ contains a value-assignment to each exogenous variable, which now includes the background variables and the error-terms. Now, a probabilistic causal model (or probabilistic SEM) is a tuple $(\mathcal{M}, \mathcal{P})$, consisting of an SEM $\mathcal{M}$ and a probability distribution $\mathcal{P}$ over the cartesian product of the value sets of all exogenous variables in $\mathcal{M}$.[27] We can, for example, represent the probabilistic forest fire scenario by a probabilistic SEM $(\mathcal{M}^F, \mathcal{P})$ as presented in figure 2.5.

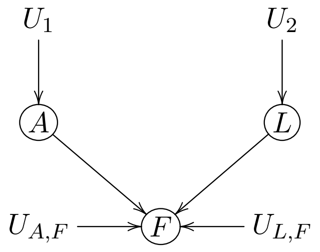
Notice that our description of $\mathcal{P}$ induces a unique joint distribution, that ascribes every possible context $\vec{u} = (u_1, u_2, u_{AF}, u_{LF})$ for $\mathcal{M}$ a probability $\mathcal{P}(\vec{u})$. In an SEM $\mathcal{M}$ the value of every endogenous variable is completely determined by the values $\vec{u}$ of the exogenous variables $\vec{U}$ in $\mathcal{M}$. It follows from this that the probability distribution $\mathcal{P}$ over the cartesian product of the value sets of the exogenous variables uniquely determines a probability distribution (which I will also denote by $\mathcal{P}$) over the cartesian product of the value sets of all variables in $\mathcal{M}$. Let $\vec{X} \subseteq \mathcal{V}$ and $\vec{x} \in \mathcal{R}(\vec{X})$, then we have:[28]

$$\mathcal{P}(\vec{X} = \vec{x}) = \sum_{\{\vec{u}_i : (\mathcal{M}, \vec{u}_i) \models \vec{X} = \vec{x}\}} \mathcal{P}(\vec{u}_i) \tag{2.4}$$

[26] Although an objective chance interpretation of $\mathcal{P}$ is generally possible, we should keep in mind that chance functions ascribe only trivial values to past events, including events of causal production. As we will see in upcoming examples, we often don't have all the information about what events and which causal productions actually took place in past scenarios. In those cases, it makes more sense to use SEMs in which the probability distributions are interpreted as credence-functions.

[27] Here again, a crucial difference between a probabilistic causal model $(\mathcal{M}, \mathcal{P})$ and a causal setting $(\mathcal{M}, \vec{u})$, which represents only deterministic causal relationships, is that in the latter case, $\mathcal{M}$ already entails all information about the causal relationships in the represented scenario, while the context $\vec{u}$ only entails information about the actual values of the variables. This is different in a probabilistic causal model $(\mathcal{M}, \mathcal{P})$ that includes error-terms. Here the probability distribution $\mathcal{P}$ does not only contain probabilistic information about the actual values of the variables, but also quantitative information about the probabilistic causal relationships in the scenario.

[28] See (Pearl, 2000, p. 205).



- $A := U_1$
- $L := U_2$
- $F := (A \wedge U_{A,F}) \vee (L \wedge U_{L,F})$

- $\mathcal{P}(U_1 = 1) = 0.05$
- $\mathcal{P}(U_2 = 1) = 0.1$
- $\mathcal{P}(U_{A,F} = 1) = 0.7$
- $\mathcal{P}(U_{L,F} = 1) = 0.7$

Figure 2.5: $(\mathcal{M}^F, \mathcal{P})$ - the probabilistic forest fire scenario.

If all exogenous variables in $\mathcal{M}$ are jointly independent, then $\mathcal{P}$ fulfills the Causal Markov Condition (CMC) relative to $\mathcal{M}$'s correspronding graph.[29]

In CBNs we have pointed out, that it is reasonable to take certain conditional probabilities, namely the specifiers, as basic and as primarily epistemically accessible. Those conditional probabilities encode information about the strength of the probabilistic causal relationships in the model. Unconditional probabilities of events, on the other hand, are understood to be derived from those conditonal probabilities. We have pointed out that this position is by now well entrenched in the phillosophical literature and is backed by authors like (Hájek, 2003), (Popper, 1959), or (Climenhaga, 2020). Now, one could find fault with the SEM approach to probabilistic causal relationships, because it seems to depart from the position that treats conditional probabilities as basic. After all, we have just shown how the distribution $\mathcal{P}$ in a probabilistic SEM is constructed from a few apperently unconditional probabilities, namely the ones explicitly stated in Figure 2.5. But notice that the probability of an error-term taking on a certain value, for example $\mathcal{P}(U_{A,F} = 1)$, is actually a conditional probability. This becomes apparent, if we bring back to mind how we interpret the term $U_{A,F} = 1$. We have pointed out that $U_{A,F} = 1$ represents a conditional, namely the fact that $A = 1$ causally produces $F = 1$, given that $A = 1$ is the case. $\mathcal{P}(U_{A,F} = 1)$ should therefore not be interpreted as an unconditional probability, but rather as a probability of an event ($A = 1$ produces $F = 1$) under the condition that another event ($A = 1$) happened. So, just like in the CBN-approach, we will consider certain conditional probabilities, namely those that encode information about the strength of probabilistic causal relationships, as basic and we will use them to derive unconditional probabilities of endogenous events in the model.

Like we have done in the CBN-approach, we can again define a formal language for probabilistic SEMs. 'Primitive events' are formulas of the form '$P(\vec{X} = \vec{x}) = p$' (read as: the probability of the event $\vec{X} = \vec{x}$ is p). Here again, 'P' is part of the object language, while '$\mathcal{P}$' is part of the meta-language. We will say $(\mathcal{M}, \mathcal{P}) \models P(\vec{X} = \vec{x}) = p$ if and only if $\sum_{\{\vec{u}_i | (\mathcal{M}, \vec{u}_i) \models \vec{X} = \vec{x}\}} \mathcal{P}(\vec{u}_i) = p$.

[29] See (Pearl, 2000, p. 30) for a proof.

Just like in deterministic SEMs, the *do*-operator for a probabilistic SEM $(\mathcal{M}, \mathcal{P})$ is defined as an operator on $\mathcal{M}$. The intervention $do(X = x)$ replaces the structural equation $\mathcal{F}_X$ in $\mathcal{M}$ with the assignment $X = x$.[30] With interventions being formally defined, we can again introduce probabilistic intervention counterfactuals of the form: 'the probability of the event ϕ would be p, if $\vec{X} = \vec{x}$ would be brought about by an intervention', formally: $[\vec{X} \leftarrow \vec{x}]P(\phi) = p$. We say $(\mathcal{M}, \mathcal{P}) \models [\vec{X} \leftarrow \vec{x}]P(\phi) = p$ if and only if $\sum_{\left\{\vec{u}_i | (\mathcal{M}_{do(\vec{X}=\vec{x})}, \vec{u}_i) \models \phi\right\}} \mathcal{P}(\vec{u}_i)$. We can again formulate more complex couterfactual formulas like: $[\vec{X} \leftarrow \vec{x}]P(\phi) > [\vec{Y} \leftarrow \vec{y}]P(\phi)$. We say $(\mathcal{M}, \mathcal{P}) \models [\vec{X} \leftarrow \vec{x}]P(\phi) > [\vec{Y} \leftarrow \vec{y}]P(\phi)$ if and only if $\sum_{\left\{\vec{u}_i | (\mathcal{M}_{do(\vec{X}=\vec{x})}, \vec{u}_i) \models \phi\right\}} \mathcal{P}(\vec{u}_i) > \sum_{\left\{\vec{u}_j | (\mathcal{M}_{do(\vec{Y}=\vec{y})}, \vec{u}_j) \models \phi\right\}} \mathcal{P}(\vec{u}_j)$.

2.4.2 SEM's, Determinism, and Indeterminism

Pearl suggests that probabilistic SEMs embody a deterministic world view. He writes:

> "In these models, causal relationships are expressed in the form of deterministic, *functional* equations, and probabilities are introduced through the assumption that certain variables in the equations are unobserved. This reflects Laplace's (1814) conception of natural phenomena, according to which nature's laws are deterministic and randomness surfaces owing merely to our ignorance of the underlying boundary conditions" (Pearl, 2000, p. 26).

I disagree with this assessment. I think that, just like CBNs, probabilistic SEMs can be used to represent all kinds of probabilistic causal relationships, genuinely indeterministic causal relationships and imprecise or incomplete descriptions of underlying deterministic relationships. The use of probabilistic SEMs is not associated with an ontological claim about determinism. As already pointed out, I agree with Pearl that many of the causal relationships, that we describe as probabilistic, are actually only incomplete or imprecise descriptions of underlying deterministic relationships. Imagine that the probabilistic causal relationship between a cause $C = 1$ and an effect $E = 1$, as described by the structural equation $E := C \wedge U_{C,E}$, does indeed result from an incomplete or imprecise description of an underlying deterministic causal relationship. Then the error-term $U_{C,E}$ represents whether the complete cause, which $C = 1$ only partially describes, is actually the case. This seems to be the interpretation of error-terms that Pearl considers to be universally valid. And it leaves no room for genuine indeterministic causal relationships. But I have proposed a broader interpretation of error-terms, according to which $U_{C,E} = 1$ represents that, given that $C = 1$ is the case, $C = 1$ successfully produces $E = 1$. This interpretation is consistent with genuine indeterministic causal relationships (since the success of the production may be up to pure chance) as well as with underlying deterministic causal relationships (since the success of the production may depend on whether the remaining, unnamed conditions of the full, deterministic cause are fulfilled).

[30] By changing the causal model $\mathcal{M}$ of a probabilistic causal model $(\mathcal{M}, \mathcal{P})$, the *do*-operator also induces a modification of the probability distribution $\mathcal{P}$. This modification is described by the semantics for intervention counterfactuals.

2.4.3 Actual Causation in Probabilistic SEMs

Now, how can we capture the concept of actual causation in probabilistic SEMs? In the CBN-approach, we faced several obstacles in trying to transfer Halpern and Pearl's definition of actual causation into contexts with probabilistic causal relationships. Fenton-Glynn's definition PC is the result of trying to overcome these obstacles. Luckily, in the SEM-approach to probabilistic causal models, we do not face the same obstacles when we aim to transfer the HP-definition of actual causation into contexts with probabilistic causal relationships. Actually, we do not face any obstacles at all, since a probabilistic SEM is nothing else than a deterministic causal model $\mathcal{M}$ and a probability distribution over all the possible contexts for $\mathcal{M}$, which amounts to a probability distribution over deterministic causal settings. Since Halpern and Pearl's definition of actual causation is designed for the framework of deterministic causal settings, we can use this very same definition in the following way: We can determine for each possible causal setting $(\mathcal{M}, \vec{u}_i)$, whether a certain candidate cause $\vec{X} = \vec{x}$ is an actual cause of the event ϕ in $(\mathcal{M}, \vec{u}_i)$, that is whether $(\mathcal{M}, \vec{u}_i) \models \vec{X} = \vec{x} \rightsquigarrow \phi$. Based on the results, we can then determine the probability of $\vec{X} = \vec{x}$ being an actual cause of ϕ in the probabilistic causal model $(\mathcal{M}, \mathcal{P})$ in the following way:[31]

Probability of being an Actual Cause. *Let $(\mathcal{M}, \mathcal{P})$ be a probabilistic causal model. The probability of $\vec{X} = \vec{x}$ being an actual cause of ϕ in $(\mathcal{M}, \mathcal{P})$ (or short: $\mathcal{P}(\vec{X} = \vec{x} \rightsquigarrow^{\mathcal{M}} \phi)$) is given by:*[32]

$$\mathcal{P}(\vec{X} = \vec{x} \rightsquigarrow^{\mathcal{M}} \phi) = \sum_{\{\vec{u}_i : (\mathcal{M}, \vec{u}_i) \models \vec{X} = \vec{x} \rightsquigarrow \phi\}} \mathcal{P}(\vec{u}_i) \tag{2.5}$$

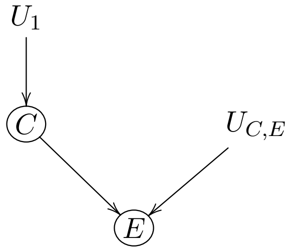
So, instead of constructing a new criterion of actual causation for probabilistic SEMs, we simply continue to use the criterion of actual causation for deterministic causal settings and provide a method to determine the probability of being such an actual cause. This approach corresponds to what I have called the de facto dependence account of actual causation. It assumes that the cause candidate either does produce the effect candidate or it does not. So, either there is a de facto dependence of the effect on the cause, or there is not. This is true irrespective of whether we are dealing with a deterministic causal relationship or with a genuinely indeterministic causal relationship. In both cases, the relevant causal production either takes place or it does not. Probabilities come into play as uncertainties about whether the causal production happens or not. These probabilities can be interpreted as objective chances, in which case only future events will have non-trivial probabilities, or as credences. To illustrate, how much the SEM-approach to actual causation differs from the CBN-approach, let us consider some further examples.

[31] This is just the approach proposed by Halpern (2016, p. 46 ff.).

[32] Just to be clear about the notation: '$\vec{X} = \vec{x} \rightsquigarrow \phi$', as introduced in the HP-definition of actual causation in section 1.3, is a formula of the object language that can be evaluated in different causal settings. '$\mathcal{P}(\vec{X} = \vec{x} \rightsquigarrow^{\mathcal{M}} \phi)$', on the other hand, is a sentence of the meta-language, which makes an assertion about the probability of an event being an actual cause of another event relative to a specific probabilistic SEM. In the following, I will omit the superscript '$\mathcal{M}$' in '$\mathcal{P}(\vec{X} = \vec{x} \rightsquigarrow^{\mathcal{M}} \phi)$', if there is no risk of confusion and as long as the context makes sufficiently clear, which causal model $\mathcal{M}$ is being discussed.

2.4.4 Determining the Probability of Being an Actual Cause

If we want to identify relations of actual causation in a CBN $(\mathcal{G}, \mathcal{P})$ that is supposed to represent a past token scenario, we should not incorporate any trumping evidence about the actual values of the variables in the model, since, as pointed out in section 2.3, this would override information about probabilistic causal relationships in the scenario, that are needed to identify relations of actual causation. This is different in the SEM-approach to actual causation, where it is all about the question of whether the cause candidate did produce (or more precisely: did contribute to the production of) the effect candidate. Trumping evidence about which events in the given scenario actually did and did not happen, are often highly useful for answering this question more reliably. Consider the following simple scenario, in which we know that an effect $E = 1$ has only one potential cause, namely $C = 1$. And we know from statistical data that there is a probability of 0.5 that $C = 1$ causally yields $E = 1$, given that $C = 1$ happens. A token of this scenario can be represented by a probabilistic SEM as shown in figure 2.6, in which we assume that $C = 1$ itself has a probability of 0.1.



- $C := U_1$
- $E := C \wedge U_{C,E}$

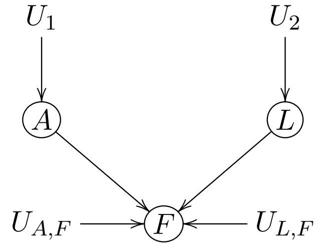
- $\mathcal{P}(U_1 = 1) = 0.1$
- $\mathcal{P}(U_{C,E} = 1) = 0.5$

Figure 2.6: Single cause with single effect.

According to this model, there is a probability of $\mathcal{P}(U_{C,E} = 1) \times \mathcal{P}(U_1 = 1) = 0.5 \times 0.1 = 0.05$ that $C = 1$ caused $E = 1$. But now imagine that, through some kind of trumping evidence like direct observation or a reliable testimony, we come to learn that $E = 1$ actually took place. This clearly influences our degree of belief in the fact that $C = 1$ caused $E = 1$. Since there is no other potential cause of $E = 1$, the information that $E = 1$ happened should rise our credence in the fact that $C = 1$ caused $E = 1$ to 1.[33] This shows, the degree of belief in whether, for example, $E = 1$ is the case is highly relevant for determining the degree of belief in whether $C = 1$ is an actual cause of $E = 1$. So, if we want to determine the probabilities of being an actual cause in an SEM $(\mathcal{M}, \mathcal{P})$ that is supposed to represent a past token scenario, we should incorporate into $\mathcal{P}$ all trumping evidence, that we currently have, about the actual values of the variables in the model. I will now present a method for constructing such a probabilistic SEM. As an example, I will use the well-known probabilistic forest fire scenario, in which $A = 1$, $L = 1$, and $F = 1$ are considered to be known to have happened, and in which $L = 1$ (as well as $A = 1$) yields $F = 1$ with a probability of 0.7, when no other causal influence on F is present.

[33] This holds under the assumption that $E = 1$ does not happen spontaneously, that is without any cause.

As we already know, the scenario can be represented by the causal model $\mathcal{M}^F$ as presented in figure 2.7.



- $A := U_1$
- $L := U_2$
- $F := (A \wedge U_{A,F}) \vee (L \wedge U_{L,F})$

Figure 2.7: $\mathcal{M}^F$ - the probabilistic forest fire scenario.

But what about the probability distribution $\mathcal{P}$? As pointed out above, the distribution $\mathcal{P}$ over the values of all the variables in the model can be generated from a distribution over the values of the exogenous variables. We can therefore start by determining the values of $\mathcal{P}(U_1 = 1)$, $\mathcal{P}(U_2 = 1)$, $\mathcal{P}(U_{A,F} = 1)$, and $\mathcal{P}(U_{L,F} = 1)$. Since we presupposed that $A = 1$ and $L = 1$ are known to have happened, we have $\mathcal{P}(U_1 = 1) = \mathcal{P}(U_2 = 1) = 1$. For the values of $\mathcal{P}(U_{A,F} = 1)$ and $\mathcal{P}(U_{L,F} = 1)$ our knowledge about the strength of the type-level probabilistic causal relationships between A and F, as well as between L and F, form the first best estimates. We presupposed that, according to statistical data, $A = 1$-type events, if present, yield $F = 1$-type events with a relative frequency of 0.7. The Principle of Direct Probability therefore recommends to ascribe an epistemic probability of 0.7 to the fact that the $A = 1$-token in question, if present, successfully yields the $F = 1$-token in question. This gives us: $\mathcal{P}(U_{A,F} = 1) = 0.7$. By the same argumentation we get: $\mathcal{P}(U_{L,F} = 1) = 0.7$.

According to formula (2.4), we can determine the probability of any event $\vec{X} = \vec{x}$ in a probabilistic SEM $(\mathcal{M}, \mathcal{P})$, if we only know the value of $\mathcal{P}(\vec{u})$ for every possible context $\vec{u}$ for $\mathcal{M}$. In our example, a context $\vec{u}$ for $\mathcal{M}^F$ is given by the tuple $\vec{u} = (u_1, u_2, u_{AF}, u_{LF})$. All contexts with either $u_1 = 0$ or $u_2 = 0$ have a probability of 0. The remaining contexts are: $\vec{u_0} = (1, 1, 1, 1), \vec{u_1} = (1, 1, 1, 0), \vec{u_2} = (1, 1, 0, 1), \vec{u_3} = (1, 1, 0, 0)$, for which we have:

- $\mathcal{P}(\vec{u_0}) = 0.7 \times 0.7 = 0.49$
- $\mathcal{P}(\vec{u_1}) = 0.7 \times 0.3 = 0.21$
- $\mathcal{P}(\vec{u_2}) = 0.3 \times 0.7 = 0.21$
- $\mathcal{P}(\vec{u_3}) = 0.3 \times 0.3 = 0.09$

Now, applying formula (2.4) gives us the following probability for $F = 1$:

$$\mathcal{P}(F = 1) = \sum_{\vec{u}_i \in \{\vec{u}_i : (\mathcal{M}^F, \vec{u}_i) \models F = 1\}} \mathcal{P}(\vec{u}_i) = \mathcal{P}(\vec{u}_0) + \mathcal{P}(\vec{u}_1) + \mathcal{P}(\vec{u}_2) = 0.91 \tag{2.6}$$

This shows that the distribution $\mathcal{P}$ does not yet adequately represent our current degrees of belief concerning the given token scenario, since we have supposed to know that $F = 1$ happened. To fix this, we have to update $\mathcal{P}$ by conditioning on $F = 1$. It is crucial to see that conditioning on $F = 1$ is not problematic in the current framework like it was for the application of the PC-definition of actual causation in the CBN-framework. There, conditioning on any trumping evidence about the actual value of an endogenous variable would have led to a loss of causal information that was relevant for the application of the PC-condition. This is not the case in the current framework. The relevant causal information that is needed for an application of the HP-definition of actual causation is stored within the structural equations of the SEM. This information is not overwritten when updating the probability distribution $\mathcal{P}$ in a probabilistic SEM by conditioning on some trumping evidence about the actual value of an endogenous variable.

Now, here is a general procedure for updating a probability distribution $\mathcal{P}_{pre}$ of a probabilistic SEM $(\mathcal{M}, \mathcal{P}_{pre})$ by conditioning on an event $\vec{X} = \vec{x}$. We assume that the values $\mathcal{P}_{pre}(\vec{u_i})$ for all contexts $\vec{u_i}$ for $\mathcal{M}$ are given. To determine the updated distribution $\mathcal{P}(\cdot) = \mathcal{P}_{pre}(\cdot|\vec{X} = \vec{x})$, it suffices to determine $\mathcal{P}(\vec{u_i})$ for all contexts $\vec{u_i}$ for $\mathcal{M}$. Now, according to the ratio formula, we have for any event $\vec{X} = \vec{x}$ in a probabilistic causal model $(\mathcal{M}, \mathcal{P}_{pre})$:

$$\mathcal{P}_{pre}(\vec{u_i} \mid \vec{X} = \vec{x}) = \frac{\mathcal{P}_{pre}(\vec{X} = \vec{x} \mid \vec{u_i}) \times \mathcal{P}_{pre}(\vec{u_i})}{\mathcal{P}_{pre}(\vec{X} = \vec{x})} \tag{2.7}$$

Now, for every context $\vec{u_i}$, we have either $(\mathcal{M}, \vec{u_i}) \models \vec{X} = \vec{x}$ or $(\mathcal{M}, \vec{u_i}) \not\models \vec{X} = \vec{x}$. For every $\vec{u_i}$ with $(\mathcal{M}, \vec{u_i}) \not\models \vec{X} = \vec{x}$, we have $\mathcal{P}_{pre}(\vec{X} = \vec{x} \mid \vec{u_i}) = 0$ and therefore $\mathcal{P}_{pre}(\vec{u_i} \mid \vec{X} = \vec{x}) = 0$. For every $\vec{u_i}$ with $(\mathcal{M}, \vec{u_i}) \models \vec{X} = \vec{x}$, we have $\mathcal{P}_{pre}(\vec{X} = \vec{x}|\vec{u_i}) = 1$ and therefore:

$$\mathcal{P}_{pre}(\vec{u_i}|\vec{X} = \vec{x}) = \frac{\mathcal{P}_{pre}(\vec{u_i})}{\mathcal{P}_{pre}(\vec{X} = \vec{x})} = \frac{\mathcal{P}_{pre}(\vec{u_i})}{\sum_{\{\vec{u}_j:(\mathcal{M},\vec{u}_j)\models\vec{X}=\vec{x}\}} \mathcal{P}_{pre}(\vec{u_j})} \tag{2.8}$$

This gives us a fairly simple procedure for updating any distribution $\mathcal{P}_{pre}$ in a given probabilistic causal model $(\mathcal{M}, \mathcal{P}_{pre})$ by some new evidence $\vec{X} = \vec{x}$ with $\vec{X}$ being endogenous variables in $\mathcal{M}$. We only have to check for all contexts $\vec{u_i}$, if $\vec{X} = \vec{x}$ holds in the causal setting $(\mathcal{M}, \vec{u_i})$. We then have:

(1) For any $\vec{u_i}$ with $(\mathcal{M}, \vec{u_i}) \not\models \vec{X} = \vec{x}$: $\mathcal{P}_{pre}(\vec{u_i} \mid \vec{X} = \vec{x}) = 0$

(2) For any $\vec{u_i}$ with $(\mathcal{M}, \vec{u_i}) \models \vec{X} = \vec{x}$: $\mathcal{P}_{pre}(\vec{u_i}|\vec{X} = \vec{x}) = \frac{\mathcal{P}(\vec{u_i})}{\sum_{\{\vec{u}_j:(\mathcal{M},\vec{u}_j)\models\vec{X}=\vec{x}\}} \mathcal{P}(\vec{u_j})}$

In our forest fire scenario, we have the following preliminary distribution:

- $\mathcal{P}_{pre}(\vec{u}_0) = 0.7 \times 0.7 = 0.49$
- $\mathcal{P}_{pre}(\vec{u}_1) = 0.7 \times 0.3 = 0.21$
- $\mathcal{P}_{pre}(\vec{u}_2) = 0.3 \times 0.7 = 0.21$
- $\mathcal{P}_{pre}(\vec{u}_3) = 0.3 \times 0.3 = 0.09$

We can now use the just explicated method for conditioning on our knowledge that $F = 1$ holds:

- $P(\vec{u_0}) = P_{pre}(\vec{u}_0|F = 1)) = \frac{P_{pre}(\vec{u_0})}{\sum_{i=0}^{2} P_{pre}(\vec{u_i})} = \frac{0.49}{0.91} = 0.538$
- $P(\vec{u_1}) = P_{pre}(\vec{u}_1|F = 1)) = \frac{P_{pre}(\vec{u_1})}{\sum_{i=0}^{2} P_{pre}(\vec{u_i})} = \frac{0.21}{0.91} = 0.231$
- $P(\vec{u_2}) = P_{pre}(\vec{u}_2|F = 1)) = \frac{P_{pre}(\vec{u_2})}{\sum_{i=0}^{2} P_{pre}(\vec{u_i})} = \frac{0.21}{0.91} = 0.231$
- $P(\vec{u_3}) = P_{pre}(\vec{u}_3|F = 1)) = 0$, since $(\mathcal{M}^F, \vec{u_3}) \not\models F = 1$.

With the updated distribution $\mathcal{P}$ we now have a probability distribution that adequately represents our current knowledge about the values of the variables in $\mathcal{M}^F$. It therefore tells us for any possible context u_i for $\mathcal{M}^F$, to which degree we actually belief, that the causal setting $(\mathcal{M}^F, \vec{u_i})$ is the correct description of the actual situation. We can now use $\mathcal{P}$ to determine $\mathcal{P}(A = 1 \rightsquigarrow F = 1)$ and $\mathcal{P}(L = 1 \rightsquigarrow F = 1)$. According to equation (2.5), we have:

$$\mathcal{P}(A = 1 \rightsquigarrow F = 1) = \sum_{\{\vec{u}_i : (\mathcal{M}^F, \vec{u}_i) \models A=1 \rightsquigarrow F=1\}} \mathcal{P}(\vec{u}_i) \tag{2.9}$$

Now, it is easy to see that $A = 1$ is an actual cause of $F = 1$ in $(\mathcal{M}^F, \vec{u_0})$ and in $(\mathcal{M}^F, \vec{u_1})$. So, we have: $\mathcal{P}(A = 1 \rightsquigarrow F = 1) = \mathcal{P}(u_0) + \mathcal{P}(u_1) = 0.769$. Analogously, we have that $L = 1$ is an actual cause of $F = 1$ in $(\mathcal{M}^F, \vec{u_0})$ and in $(\mathcal{M}^F, \vec{u_2})$. We therefore have: $\mathcal{P}(L = 1 \rightsquigarrow F = 1) = \mathcal{P}(u_0) + \mathcal{P}(u_2) = 0.769$. So, learning that $F = 1$ actually happened slightly increases our degree of belief that $A = 1$ is an actual cause of $F = 1$ and our degree of belief that $L = 1$ is an actual cause of $F = 1$.

The following five steps summarize the procedure for (1) constructing a probabilistic SEM $(\mathcal{M}, \mathcal{P})$ that adequately represents a rational agent's credences about a past token scenario and for (2) determining the probability of $\vec{X} = \vec{x}$ being an actual cause of ϕ in $(\mathcal{M}, \mathcal{P})$:[34]

PROBAC. *For determining the probability of an event $\vec{X} = \vec{x}$ being an actual cause of an event ϕ in a past token scenario whose causal relationships are captured by the SEM $\mathcal{M}$, follow the following instructions:*

(1) Create a preliminary probability distribution $\mathcal{P}_{pre}$ such that: (a) $\mathcal{P}_{pre}$ ascribes probabilities to the background variables in $\mathcal{M}$ that are consistent with the current knowledge about the token scenario and (b) if $U_{X,Y}$ is an error-term that represents, whether $X = x$ causally produces $Y = y$, given $X = x$, then $\mathcal{P}_{pre}(U_{X,Y} = 1)$ should be equal to the statistically determined relative frequency with which $X = x$-type events, if present, successfully produce $Y = y$-type events.[35]

(2) Determine the values $\mathcal{P}_{pre}(\vec{u}_i)$ for all contexts $\vec{u}_i$ for $\mathcal{M}$.

(3) In case of additional knowledge about the actual values $\vec{z}$ of endogenous variables $\vec{Z}$ in $\mathcal{M}$, update $\mathcal{P}_{pre}$ by conditioning on $\vec{Z} = \vec{z}$ to obtain $\mathcal{P} = \mathcal{P}_{pre}(\cdot \mid \vec{Z} = \vec{z})$

[34] The acronym 'PROBAC' simply stands for 'probability of being an actual cause'.

[35] As pointed out before, we will deal with the question of how to determine this relative frequency in an upcoming chapter.

(4) Test for each possible causal setting $(\mathcal{M}, \vec{u}_i)$, if $\vec{X} = \vec{x}$ is an actual cause of ϕ in $(\mathcal{M}, \vec{u}_i)$.

(5) Calculate the sum of the probabilities of all contexts $\vec{u}_i$ with $\vec{X} = \vec{x}$ being an actual cause of ϕ in $(\mathcal{M}, \vec{u}_i)$: $\sum_{\{\vec{u}_i:(\mathcal{M},\vec{u}_i)\models\vec{X}=\vec{x}\rightsquigarrow\phi\}} \mathcal{P}(\vec{u}_i)$.

2.4.5 EIC-Counterfactuals

Steps 3 to 5 of PROBAC are reminiscent of a method that Pearl (2000, p. 206) has put forward to evaluate a certain kind of counterfactual that has the following form: 'Given that $\vec{Z} = \vec{z}$, the event ϕ would have probability p, if we would intervene to set $\vec{X}$ to $\vec{x}$'. I will call such a counterfactual an *EIC-counterfactual* to differentiate them from intervention counterfactuals of the form $[\vec{X} \leftarrow \vec{x}]P(\phi) = p$, which simply says: 'If we would intervene to set $\vec{X}$ to $\vec{x}$, then ϕ would have probability p'.[36] EIC-counterfactuals will have the following form in our object language: $[\vec{X} \leftarrow \vec{x}]P(\phi|\vec{Z} = \vec{z}) = p$. To evaluate $[\vec{X} \leftarrow \vec{x}]P(\phi|\vec{Z} = \vec{z}) = p$ in a probabilistic causal model $(\mathcal{M}, \mathcal{P})$, we first have to update the distribution $\mathcal{P}$ by conditioning on $\vec{Z} = \vec{z}$ to obtain $\mathcal{P}' = \mathcal{P}(\cdot \mid \vec{Z} = \vec{z})$ (this corresponds to step 3 in PROBAC). We then intervene on $\mathcal{M}$ to obtain $\mathcal{M}_{do(\vec{X}=\vec{x})}$ (this stands in analogy to step 4 in PROBAC). Finally, we calculate the sum of the probabilities of all contexts $\vec{u}_i$ with $(\mathcal{M}_{do(\vec{X}=\vec{x})}, \vec{u}_i) \models \phi$ (this corresponds to step 5 in PROBAC).

Now, while EIC-counterfactuals can be evaluated in probabilistic SEMs through this three-step procedure, it seems that EIC-counterfactuals of the form '$[\vec{X} \leftarrow \vec{x}]P(\phi|\vec{Z} = \vec{z}) = p$' cannot be evaluated in CBNs due to the fact that updating the distribution $\mathcal{P}$ in a CBN $(\mathcal{G}, \mathcal{P})$ by conditioning on $\vec{Z} = \vec{z}$ overwrites the relevant causal information that we need to evaluate the embedded counterfactual '$[\vec{X} \leftarrow \vec{x}]P(\phi) = p$'. This argument has been put forward by Pearl in (Pearl, 1999) and (Pearl, 2000, pp. 26 ff.), which is why he ascribes SEMs a higher expressiveness than CBNs when it comes to causally relevant intervention counterfactuals. But with a slight modification of the standard CBN-framework, one could argue that EIC-counterfactuals can indeed be evaluated in CBN's. We can employ, as proposed above, two distinct distributions over the cartesian product of the value sets of the variables in a CBN: The first, denoted by $\mathcal{P}$, encodes the information about the strengths of the probabilistic causal relationships in the scenario. The second, denoted by $\mathcal{C}_{pr}$, represents our current degrees of beliefs about the actual values of the variables in the model. We can then evaluate an ECI-counterfactual $[\vec{X} \leftarrow \vec{x}]P(\phi|\vec{Z} = \vec{z}) = p$ in the following way: We first update the probability distribution $\mathcal{C}_{pr}$ by conditioning on $\vec{Z} = \vec{z}$. We then evaluate the embedded intervention-counterfactual $[\vec{X} \leftarrow \vec{x}]P(\phi) = p$ by checking whether $(\mathcal{G}, \mathcal{P}) \models [\vec{X} = \vec{x}]P(\phi) = p$. According to this method, the condition $\vec{Z} = \vec{z}$ has obviously no impact at all on the evaluation of the embedded counterfactual $[\vec{X} \leftarrow \vec{x}]P(\phi) = p$, since the embedded counterfactual is always evaluated relative to the causal distribution $\mathcal{P}$, which is not changed by the assumption that $\vec{Z} = \vec{z}$ holds. For a given CBN $(\mathcal{G}, \mathcal{P})$, it therefore holds that for any $\vec{Z} = \vec{z}$, $[\vec{X} \leftarrow \vec{x}]P(\phi|\vec{Z} = \vec{z}) = p$ is true in $(\mathcal{G}, \mathcal{P})$ if and only if $[\vec{X} \leftarrow \vec{x}]P(\phi) = p$ is true in $(\mathcal{G}, \mathcal{P})$. The same does clearly not hold for SEMs. So, according to this proposal, CBNs are

[36] The acronym 'EIC' stands for 'embedded in an indicative conditional'. Pearl (2000) calls sentences of the form $[\vec{X} \leftarrow \vec{x}]P(\phi) = p$ *interventions* and what I call EIC-counterfactuals, he simply calls *counterfactuals*. I deviate from this terminology to avoid a confusion with the intervention-operator.

indeed able to evaluate EIC-counterfactuals. They only do so very differently than probabilistic SEMs.

The two different approaches to evaluating EIC-counterfactuals perfectly reflects the frameworks' different interpretations of actual causation. To illustrate, consider again the probabilistic forest fire scenario with the probabilistic causal relationships as presented in the previous section. Imagine that the scenario is a past token scenario and we already know that $A = 0$ and $L = 1$, but we do not know the actual value of F. We are interested in the following EIC-counterfactual: $[A \leftarrow 1]P(F = 1|F = 0) = p$.

In the SEM-framework, we start with the the causal model $\mathcal{M}^F$, as depicted in figure 2.7, and with the following distribution:

- $\mathcal{P}(U_1 = 1) = 0$
- $\mathcal{P}(U_2 = 1) = 1$
- $\mathcal{P}(U_{A,F} = 1) = 0.7$
- $\mathcal{P}(U_{L,F} = 1) = 0.7$

$\mathcal{P}$ ascribes a probability of 0 to all contexts for $\mathcal{M}^F$, except for $\vec{u_0} = (0, 1, 1, 1), \vec{u_1} = (0, 1, 1, 0), \vec{u_2} = (0, 1, 0, 1), \vec{u_3} = (0, 1, 0, 0)$, for which we have:

- $\mathcal{P}(\vec{u}_0) = 0.7 \times 0.7 = 0.49$
- $\mathcal{P}(\vec{u}_1) = 0.7 \times 0.3 = 0.21$
- $\mathcal{P}(\vec{u}_2) = 0.3 \times 0.7 = 0.21$
- $\mathcal{P}(\vec{u}_3) = 0.3 \times 0.3 = 0.09$

We now have to update $\mathcal{P}$ by conditioning on $F = 0$. The updated distribution $\mathcal{P}'(\cdot) = \mathcal{P}(\cdot|F = 0)$ ascribes 0 to all contexts for $\mathcal{M}^F$, except:

- $\mathcal{P}'(\vec{u}_1) = \mathcal{P}(\vec{u}_1|F = 0)) = \frac{\mathcal{P}(\vec{u_1})}{\sum_i \mathcal{P}(\vec{u_i})} = \frac{0.21}{0.3} = 0.7$
- $\mathcal{P}'(\vec{u}_3) = \mathcal{P}(\vec{u}_3|F = 0)) = \frac{\mathcal{P}(\vec{u_3})}{\sum_i \mathcal{P}(\vec{u_i})} = \frac{0.09}{0.3} = 0.3$

We can now use the updated distribution $\mathcal{P}'$ to determine the following sum:

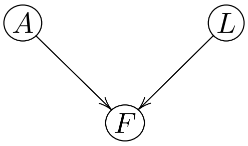
$$\sum_{\{\vec{u}_i:(\mathcal{M}^F_{do(A=1)},\vec{u}_i)\models F=1\}} \mathcal{P}'(\vec{u}_i) \tag{2.10}$$

Since $\vec{u}_1$ is the only context with $(\mathcal{M}^F_{do(A=1)}, \vec{u}_1) \models F = 1$, we have:

$$\sum_{\{\vec{u}_i:(\mathcal{M}^F_{do(A=1)},\vec{u}_i)\models F=1\}} \mathcal{P}'(\vec{u}_i) = 0.7 \tag{2.11}$$

Therefore: $(\mathcal{M}^F, \mathcal{P}) \models [A \leftarrow 1]P(F = 1|F = 0) = 0.7$.

In the CBN-framework, the CBN $(\mathcal{G}, \mathcal{P})$ as shown in figure 2.8 represents the given scenario.



- $\mathcal{P}(A=1)=0$
- $\mathcal{P}(L=1)=1$

- $\mathcal{P}(F=1|A=1,L=1)=0.91$
- $\mathcal{P}(F=1|A=1,L=0)=0.7$
- $\mathcal{P}(F=1|A=0,L=1)=0.7$
- $\mathcal{P}(F=1|A=0,L=0)=0$

Figure 2.8: CBN for the probabilistic forest fire scenario.

As pointed out above, $[A \leftarrow 1]P(F=1|F=0)=p$ is true in $(\mathcal{G},\mathcal{P})$ if and only if $[A \leftarrow 1]P(F=1)=p$ is true in $(\mathcal{G},\mathcal{P})$. To check for which value of p $[A \leftarrow 1]P(F=1|F=0)=p$ is true in $(\mathcal{G},\mathcal{P})$, we can therefore simply check for which value of p $[A \leftarrow 1]P(F=1)=p$ is true in $(\mathcal{G},\mathcal{P})$, which is just the probability of $F=1$ in $(\mathcal{G},\mathcal{P}_{do(A=1)})$, namely 0.91. This gives us: $(\mathcal{G},\mathcal{P}) \models [A \leftarrow 1]P(F=1|F=0)=0.91$.

The CBN- and the SEM-framework differ in their assessment of the given EIC-counterfactual, because they keep different things fixed, when implementing the conditions mentioned in the two antecedents of the EIC-counterfactual. The SEM-framework keeps the fact fixed that $L=1$ did not causally produce $F=1$ in the given token scenario. Since this is all that matters for a token causal relationship in the SEM-framework, $L=1$ does therefore not have any causal influence on $F=1$ in the given situation. So, in the counterfactual scenario, in which $A=1$ would have been realized by an intervention and would therefore hold additionally to $L=1$, we would only have the causal influence of $A=1$ on $F=1$, such that the probability of $F=1$ would be 0.7. The CBN-framework, on the other hand, keeps the fact fixed that $L=1$ increased the probability of $F=1$ to 0.7 in the given situation. The fact that $L=1$ did ultimately not causally produce the event $F=1$ in the given token scenario is irrelevant in the CBN-framework. What counts, is only that $L=1$ produced a probability increase for $F=1$. So, in the counterfactual scenario, in which $A=1$ would have been realized by an intervention and would therefore hold additionally to $L=1$, we would have both causal influences of $A=1$ and $L=1$ on $F=1$, such that the probability of $F=1$ would be 0.91.

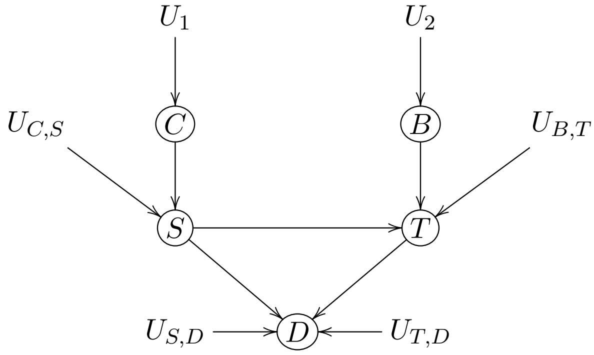
This illustrates that both frameworks have a crucially different understanding of what kind of facts or events ground relations of actual causation. For the SEM-approach, it is all about the causal production of token events. For the CBN-approach, it is all about the production of certain probability changes.

2.5 Some Further Examples

To provide some further comparisons between the SEM- and the CBN-approach to actual causation, let us consider two additional examples.

The Mafia Scenario in the SEM-approach

Let us first see how the SEM-approach deals with Fenton-Glynn's mafia example. The causal model $\mathcal{M}^M$ as shown in figure 2.9 represents the causal relationships in the scenario qualitatively.



- $C := U_1$
- $B := U_2$

- $S := C \wedge U_{C,S}$
- $T := (B \wedge U_{B,T}) \wedge \neg S$
- $D := (S \wedge U_{S,D}) \vee (T \wedge U_{T,D})$

Figure 2.9: $\mathcal{M}^M$ - the mafia scenario.

Due to our knowledge about the causal relations, we have the following preliminary probabilities: $\mathcal{P}_{pre}(U_{C,S} = 1) = 0.9$, $\mathcal{P}_{pre}(U_{B,T} = 1) = 0.9$, $\mathcal{P}_{pre}(U_{S,D} = 1) = 0.5$ and $\mathcal{P}_{pre}(U_{T,D} = 1) = 0.9$. We also have the following evidence about the actual situation: $C = 1$, $B = 1$, $S = 1$, $T = 0$, and $D = 1$. For our preliminary distribution, this means: $\mathcal{P}_{pre}(U_1 = 1) = \mathcal{P}_{pre}(U_2 = 1) = 1$. For the calculation of the updated probability distribution $\mathcal{P}(\cdot) = \mathcal{P}_{pre}(\cdot|S = 1, T = 0, D = 1)$, we only have to consider those contexts $\vec{u_i}$, for which the given evidence is true in $(\mathcal{M}^M, \vec{u_i})$, since we already know that $\mathcal{P}(\vec{u_j}) = \mathcal{P}_{pre}(\vec{u_j}|S = 1, T = 0, D = 1) = 0$ for all $\vec{u_j}$ with $(\mathcal{M}^M, \vec{u_j}) \nvDash S = 1 \wedge T = 0 \wedge D = 1$. A context is given by the tuple $\vec{u} = (u_1, u_2, u_{CS}, u_{BT}, u_{SD}, u_{TD})$. This gives us the following set of possible contexts with their preliminary probabilities:

- $\vec{u}_0 = (1,1,1,1,1,1)$ with $\mathcal{P}_{pre}(\vec{u}_0) = 1 \times 1 \times 0.9 \times 0.9 \times 0.5 \times 0.9 \approx 0.365$
- $\vec{u}_1 = (1,1,1,0,1,1)$ with $\mathcal{P}_{pre}(\vec{u}_1) = 1 \times 1 \times 0.9 \times 0.1 \times 0.5 \times 0.9 \approx 0.041$
- $\vec{u}_2 = (1,1,1,1,1,0)$ with $\mathcal{P}_{pre}(\vec{u}_2) = 1 \times 1 \times 0.9 \times 0.9 \times 0.5 \times 0.1 \approx 0.041$
- $\vec{u}_3 = (1,1,1,0,1,0)$ with $\mathcal{P}_{pre}(\vec{u}_3) = 1 \times 1 \times 0.9 \times 0.1 \times 0.5 \times 0.1 \approx 0.005$

We can now determine the updated distribution $\mathcal{P}$ by determining the values of $\mathcal{P}(\vec{u}_i) = \mathcal{P}_{pre}(\vec{u} \mid S = 1 \wedge T = 0 \wedge D = 1)$ for each of the four contexts above. According to equation (2.8), we have:

- $\mathcal{P}(\vec{u_0}) = \frac{\mathcal{P}_{pre}(\vec{u_0})}{\sum_{i=0}^{3} \mathcal{P}_{pre}(\vec{u_i})} \approx \frac{0.365}{0.452} \approx 0.807$
- $\mathcal{P}(\vec{u_1}) = \frac{\mathcal{P}_{pre}(\vec{u_1})}{\sum_{i=0}^{3} \mathcal{P}_{pre}(\vec{u_i})} \approx \frac{0.041}{0.452} \approx 0.091$
- $\mathcal{P}(\vec{u_2}) = \frac{\mathcal{P}_{pre}(\vec{u_2})}{\sum_{i=0}^{3} \mathcal{P}_{pre}(\vec{u_i})} \approx \frac{0.041}{0.452} \approx 0.091$
- $\mathcal{P}(\vec{u_3}) = \frac{\mathcal{P}_{pre}(\vec{u_2})}{\sum_{i=0}^{3} \mathcal{P}_{pre}(\vec{u_i})} \approx \frac{0.005}{0.452} \approx 0.011$

The probability distribution $\mathcal{P}$ now adequately represents our current epistemic state about the given situation. We can now come to step 4 of PROBAC. We have to check in which of the four contexts $C = 1$ is an actual cause of $D = 1$. It is easy to verify that this is the case in all four contexts $\vec{u_0}$, $\vec{u_1}$, $\vec{u_2}$, $\vec{u_3}$.[37] Proceeding with step 5, we get: $\mathcal{P}(C = 1 \rightsquigarrow D = 1) = \mathcal{P}(\vec{u_0}) + \mathcal{P}(\vec{u_1}) + \mathcal{P}(\vec{u_2}) + \mathcal{P}(\vec{u_3}) = 1$. On the other hand, in none of the four contexts with probability higher than 0 it holds that $B = 1$ is an actual cause of $D = 1$. So, we have: $\mathcal{P}(B = 1 \rightsquigarrow D = 1) = 0$.

The mafia-scenario shows, that the assessments about actual causation in the SEM-approach sometimes coincide with the assessments of the CBN-approach. Both accounts conclude in the given example, that $C = 1$ is an actual cause of $D = 1$, while $B = 1$ is no actual cause of $D = 1$. But, as the following example shows, we can generally not count on such a consensus.

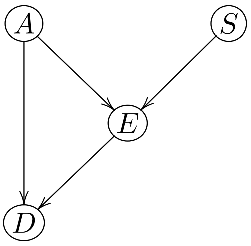
Snake Poison and the Antidote

We have already dealt with an example, namely the probabilistic forest fire scenario, in which Fenton-Glynn's PC-definition of actual causation identifies an event $\vec{X} = \vec{x}$ as an actual cause of an event ϕ, even though the SEM-approach concludes that there is a non-zero probability that $\vec{X} = \vec{x}$ is no actual cause of ϕ. I will now present an example, in which the PC-definition concludes that a given event $\vec{X} = \vec{x}$ is no an actual cause of a given event ϕ, while the SEM-approach concludes that there is a certain probability that $\vec{X} = \vec{x}$ is an actual cause of ϕ.

Imagine that a patient has been bitten by a poisonous snake and it is known from statistical data that a poisoning of this kind leads to a death of the bitten person in 50% of the cases. But luckily, we have a doctor nearby, who has the antidote to the snake poison. But this antidote is quite hazardous itself. Even though it definetely counters the impact of the snake poison in the patients body, it has the annoying side-effect that it kills half of the patients that take it through a different chemical process. For this reason, the doctor, despite always having the antidote in his medicine cabinet, has made the categorical decision, to never give any patient the antidote, no matter how many snake bites he or she has. He probably wants to be able to blame the snake and not himself for the death of the patient. Now, in the actual situation that we aim to model, the doctor follows his own provision and does not give the bitten patient the antidote. Unfortunately, the patient dies. Figure 2.10 shows the CBN $(\mathcal{G}, \mathcal{P})$ which represents the given scenario, where variable S represents whether the given patient is bitten by a snake, A

[37] In $(\mathcal{M}^M, \vec{u_1})$, $(\mathcal{M}^M, \vec{u_2})$, $(\mathcal{M}^M, \vec{u_3})$ we have a simple counterfactual dependence of $D = 1$ on $C = 1$. In $(\mathcal{M}^M, \vec{u_0})$ we have a counterfactual dependence of $D = 1$ on $C = 1$ in the contingency, in which T is kept fixed at value 0.

represents whether the doctor administers the antidote, E represents whether the snake poison unfolds its effect in the patients body and D represents whether the patient dies.



- $\mathcal{P}(A = 1) = 0$
- $\mathcal{P}(S = 1) = 1$
- $\mathcal{P}(E = 1 \mid A = 1, S = 1) = 0$
- $\mathcal{P}(E = 1 \mid A = 0, S = 1) = 1$
- $\mathcal{P}(E = 1 \mid A = 1, S = 0) = 0$
- $\mathcal{P}(E = 1 \mid A = 0, S = 0) = 0$
- $\mathcal{P}(D = 1 \mid A = 1, E = 1) = 0.75$
- $\mathcal{P}(D = 1 \mid A = 0, E = 1) = 0.5$
- $\mathcal{P}(D = 1 \mid A = 1, E = 0) = 0.5$
- $\mathcal{P}(D = 1 \mid A = 0, E = 0) = 0$

Figure 2.10: CBN for the snake scenario.

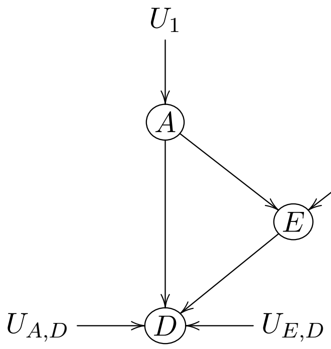
Now, according to the PC-definition of actual causation, $A = 0$ is no actual cause of $D = 1$ in $(\mathcal{G}, \mathcal{P})$.

Proof. PC1 is fulfilled, since $A = 0$ and $D = 1$ are assumed to have actually happened in our scenario. PC3 is fulfilled if PC2 is fulfilled, since $A = 0$ is a singleton. So, it all depends on PC2. So, is there any contingency in which setting A to 0 gives a higher value for $\mathcal{P}(D = 1)$ than setting A to 1? This is not the case in the actual situation. With $\vec{W} = \varnothing$, we have $\mathcal{P}_{do(A=0,\vec{Z'}=\vec{z}^*)}(D = 1) = 0.5$ for all $\vec{Z'} \subseteq \vec{Z}$. But, we also have: $\mathcal{P}_{do(A=1)}(D = 1) = 0.5$. So, condition PC2 is not fulfilled for $\vec{W} = \varnothing$. Now, the only hopeful candidate for $\vec{W}$ is $\{S, E\}$ with $S = 0$ and $E = 0$. But, it is clear, that we have $\mathcal{P}_{do(A=0,S=0,E=0,\vec{Z'}=\vec{z}^*)}(D = 1) = 0$ for $\vec{Z'} = \varnothing$. But we have $\mathcal{P}_{do(A=1,S=0,E=0)}(D = 1) = 0.5$. So, again condition PC2 is not fulfilled. □

If we model the same situation using the SEM-framework, we get the causal model $\mathcal{M}^{Sn}$ as shown in figure 2.11.

A context is given by the tuple $\vec{u} = (u_1, u_2, u_{A,D}, u_{E,D})$. According to the probabilistic causal relationships, we have $\mathcal{P}_{Pre}(U_{A,D} = 1) = 0.5$ and $\mathcal{P}_{Pre}(U_{E,D} = 1) = 0.5$. Since $A = 0$ and $S = 1$ are known to have happened, we have $\mathcal{P}_{Pre}(U_1 = 1) = 0$ and $\mathcal{P}_{Pre}(U_2 = 1) = 1$. To determine the updated distribution $\mathcal{P}(\cdot) = \mathcal{P}_{Pre}(\cdot \mid D = 1)$, we have to consider all contexts $\vec{u}_i$ with $(\mathcal{M}^{Sn}, \vec{u}_i) \models A = 0 \land S = 1 \land D = 1$. This is the case for the following contexts: $\vec{u_0} = (0, 1, 1, 1)$ and $\vec{u_1} = (0, 1, 0, 1)$, which have the following preliminary probabilities:

- $\mathcal{P}_{pre}(\vec{u}_0) = \mathcal{P}_{Pre}(u_{A,D}) \times \mathcal{P}_{Pre}(u_{E,D}) = 0.5 \times 0.5 = 0.25$
- $\mathcal{P}_{pre}(\vec{u}_1) = (1 - \mathcal{P}_{Pre}(u_{A,D})) \times \mathcal{P}_{Pre}(u_{E,D}) = 0.5 \times 0.5 = 0.25$



- $A := U_1$
- $S := U_2$
- $E := S \land \neg A$
- $D := (A \land U_{A,D}) \lor (E \land U_{E,D})$

Figure 2.11: $\mathcal{M}^{Sn}$ - the snake scenario.

The updated probability distribution $\mathcal{P}(\cdot) = \mathcal{P}_{Pre}(\cdot \mid D = 1)$ for each of these contexts is:

- $\mathcal{P}(\vec{u_0}) = \frac{\mathcal{P}_{pre}(\vec{u_0})}{\sum_{i=0}^{1} \mathcal{P}_{pre}(\vec{u_i})} = \frac{0.25}{0.5} = 0.5$
- $\mathcal{P}(\vec{u_1}) = \frac{\mathcal{P}_{pre}(\vec{u_1})}{\sum_{i=0}^{1} \mathcal{P}_{pre}(\vec{u_i})} = \frac{0.25}{0.5} = 0.5$

For determining $\mathcal{P}(A = 0 \rightsquigarrow D = 1)$ we have to check whether $A = 0$ is an actual cause of $D = 1$ in $(\mathcal{M}^{Sn}, \vec{u_0})$ and whether $A = 0$ is an actual cause of $D = 1$ in $(\mathcal{M}^{Sn}, \vec{u_1})$. First, it is obvious that in $(\mathcal{M}^{Sn}, \vec{u_0})$ the patient would still have died in any contingency, if the antidote would have been given to him. This is because $(\mathcal{M}^{Sn}, \vec{u_0}) \models U_{A,D} = 1$. Therefore, $A = 0$ is no actual cause of $D = 1$ in $(\mathcal{M}^{Sn}, \vec{u_0})$. But in $(\mathcal{M}^{Sn}, \vec{u_1})$, we have $U_{A,D} = 0$. So, setting A to 1 would not have produced $D = 1$. Quite the contrary, $A = 1$ would have caused $E = 0$, such that $(\mathcal{M}^{Sn}_{do(A=1)}, \vec{u_1}) \models D = 0$. So, if we choose $\vec{W} = \varnothing$ for the HP-criterion of actual causation, then it is easy to check that $A = 0$ passes all criteria of being an actual cause of $D = 1$ in $(\mathcal{M}^{Sn}, \vec{u_1})$. Applying step 5 of PROBAC, we get: $\mathcal{P}(A = 0 \rightsquigarrow D = 1) = \mathcal{P}(\vec{u_1}) = 0.5$.

One can nicely illustrate the divergence of both accounts of actual causation in the given example, by considering how both accounts answer the following question: In hindsight, that is given our knowledge that the snake has bitten the patient, the antidote has not been administered, and the patient died, would it have been a better decision to administer the antidote? According to the SEM-approach the answer is yes. The reason is that the SEM-approach holds the fact fixed that the snake poison successfully killed the patient in the actual situation, when evaluating the counterfactual in the question above. So, in hindsight, given the fact that the snake poison indeed killed the patient, the decision not to administer the antidote has been a certain death sentence for the patient. Would the doctor have decided to administer the antidote, there would, according to our current state of knowledge, have been a possibility that the patient would have survived. This is why, in hindsight, with the knowledge that we currently have, it would have been the better decision.[38]

[38] This does not mean, though, that the doctor is to blame for his decision. Before the patient died, he did not

According to the CBN-approach, on the other hand, the answer to the question above is no. The reason is that the CBN-approach does not hold the fact fixed that the snake poison successfully killed the patient in the actual situation, when evaluating the counterfactual in the question. Instead, the CBN-approach holds fixed the fact that the snake poison produced a 50% chance for the patient's death. But with this assumption kept fixed, both alternative actions, setting A to 1 or setting A to 0, have just the same effect on the probability of $D = 1$.

2.6 Summary and Conclusion

In this chapter, I have presented two different formal approaches to define actual causation in the context of probabilistic causal relationships. I have shown that both accounts are based on very different ideas of what actual causation really is. According to the CBN-approach, actual causation is essentially de facto probability raising, while according to the SEM-approach, actual causation is essentially a de facto dependence between events. I have also shown that, while both approaches sometimes coincide in their assessments of actual causation, there are cases in which the CBN-approach to actual causation concludes that an event $\vec{X} = \vec{x}$ is an actual cause of another event ϕ, even though the SEM-approach concludes that there is a certain non-zero probability that $\vec{X} = \vec{x}$ is no actual cause of ϕ. And there are cases in which the CBN-approach concludes that an event $\vec{X} = \vec{x}$ is no actual cause of another event ϕ, even though the SEM-approach concludes that there is a certain probability that $\vec{X} = \vec{x}$ is an actual cause of ϕ.

Now, the question is: which approach is the better one? Which framework should I use for the project of explicating causal explanations in the context of probabilistic causal relationships? There is no reason to think that one of the two approaches is outright wrong or inconsistent. Nonetheless, in the following I will build my account of causal explanation in the context of probabilistic causal relationships on the SEM-approach. One reason is simply personal intuition. I consider the SEM-account's conception of actual causation to be more intuitive than the CBN-account's conception. But I also think that this intuition is not just my own. There are good reasons to believe that the SEM-account is more in accord with common intuitions about causation than the CBN-account. The reason is that the SEM-approach takes the idea of causal productions of token events, instead of causal productions of probabilities of token events, as being essential to causation. As we will see in chapter 4, this is in accord with an empirically well founded psychological theory of how humans actually reason about probabilistic causal relationships.

know that the snake poison would ultimately succeed in killing the patient. So, with his knowledge at the time, both actions, setting A to 0 and setting A to 1, appeared equally bad to him. Only after his action $A = 0$ and after the death of the patient, we reached an epistemically better position. We are not in a better position to help the patient. That ship has sailed. But we are in a better position when it comes to our knowledge about the actual situation. We now know that the snake poison succeeded in killing the patient and with that additional knowledge, we can say that, in hindsight, it would have been better to administer the antidote.

Chapter 3

Causal Explanations in Probabilistic Contexts

3.1 Introduction

Having provided a framework for reasoning about actual causation in scenarios with probabilistic causal relationships, it is now time to clarify how to reason about causal explanations in scenarios with probabilistic causal relationships. The present chapter will transfer the definitions of causal explanation, as given in chapter 1, into the framework of probabilistic SEMs.

3.2 Strong Actual Causation in Probabilistic SEMs

Since our definitions of explanation for deterministic contexts are built on the concept of strong actual causation, a straightforward first step for transferring the definitions into the framework of probabilistic SEMs is to transfer the concept of strong actual causation into the framework of probabilistic SEMs. Luckily, this works just as easy as for the concept of actual causation. A probabilistic causal model is nothing else than a deterministic causal model $\mathcal{M}$ and a probability distribution over all the possible contexts for $\mathcal{M}$, which amounts to a probability distribution over deterministic causal settings. Since our definition of strong actual causation is designed for deterministic causal settings, we can simply continue to use the very same definition. Given a probabilistic causal model $(\mathcal{M}, \mathcal{P})$, we determine for each causal setting $(\mathcal{M}, \vec{u}_i)$ with $\mathcal{P}(\vec{u}_i) > 0$, whether the cause candidate $\vec{X} = \vec{x}$ is a strong actual cause of the event ϕ in $(\mathcal{M}, \vec{u}_i)$. Based on the results, we can then determine the probability of $\vec{X} = \vec{x}$ being a strong actual cause of ϕ in $(\mathcal{M}, \mathcal{P})$ in the following way:

Probability of Being a Strong Actual Cause. *Let $(\mathcal{M}, \mathcal{P})$ be a probabilistic causal model. The probability of $\vec{X} = \vec{x}$ being a strong actual cause of ϕ in $(\mathcal{M}, \mathcal{P})$ (or short: $\mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} \phi)$) is given by:*[1]

$$\mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} \phi) = \sum_{\{\vec{u}_i : (\mathcal{M}, \vec{u}_i) \models \vec{X} = \vec{x} \rightarrowtail \phi\}} \mathcal{P}(\vec{u}_i) \tag{3.1}$$

[1] In the following, I will omit the superscript '$\mathcal{M}$' in '$\mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} \phi)$', if there is no risk of confusion and as long as the context makes sufficiently clear, which causal model $\mathcal{M}$ is being discussed.

In analogy to PROBAC, we can use the following procedure for (1) constructing a probabilistic SEM $(\mathcal{M},\mathcal{P})$ that adequately represents a rational agent's credences about a past token scenario and for (2) determining the probability of $\vec{X}=\vec{x}$ being a strong actual cause of ϕ in $(\mathcal{M},\mathcal{P})$:[2]

PROSAC. *For determining the probability of an event $\vec{X}=\vec{x}$ being a strong actual cause of an event ϕ in a past token scenario whose causal relationships are captured by the SEM $\mathcal{M}$, follow the following instructions:*

(1) Create a preliminary probability distribution $\mathcal{P}_{pre}$ such that: (a) $\mathcal{P}_{pre}$ ascribes probabilities to the background variables in $\mathcal{M}$ that are consistent with the current knowledge about the token scenario and (b) if $U_{X,Y}$ is an error-term that represents, whether $X=x$ causally produces $Y=y$, given $X=x$, then $\mathcal{P}_{pre}(U_{X,Y}=1)$ should be equal to the statistically determined relative frequency with which $X=x$-type events, if present, successfully produce $Y=y$-type events.

(2) Determine the values $\mathcal{P}_{pre}(\vec{u}_i)$ for all contexts $\vec{u}_i$ for $\mathcal{M}$.

(3) In case of additional knowledge about the actual values $\vec{z}$ of endogenous variables $\vec{Z}$ in $\mathcal{M}$, update $\mathcal{P}_{pre}$ by conditioning on $\vec{Z}=\vec{z}$ to obtain $\mathcal{P}=\mathcal{P}_{pre}(\cdot \mid \vec{Z}=\vec{z})$.

(4) Test for each possible causal setting $(\mathcal{M},\vec{u}_i)$, if $\vec{X}=\vec{x}$ is a strong actual cause of ϕ in $(\mathcal{M},\vec{u}_i)$.

(5) Determine the sum of the probabilities of all contexts $\vec{u}_i$ with $\vec{X}=\vec{x}$ being a strong actual cause of ϕ in $(\mathcal{M},\vec{u}_i)$: $\sum_{\{\vec{u}_i:(\mathcal{M},\vec{u}_i)\models\vec{X}=\vec{x}\rightarrowtail\phi\}}\mathcal{P}(\vec{u}_i)$.

3.3 Sufficient Causes in Probabilistic SEMs

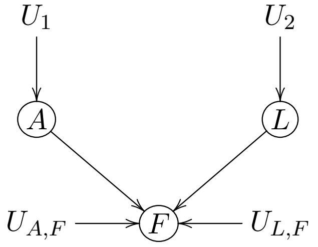
In section 1.5.7, I briefly raised the proposal to use the concept of sufficient causation in our amended definitions of explanation. I have pointed out that the resulting definitions would deal with the problematic examples from chapter 1 just as well as the amended definitions that are based on the concept of strong actual causation. We are now in a position to illustrate, why strong actual causation is nonetheless the more reasonable choice for defining concepts of explanation. The reason is that the definition of sufficient causation or, to be exact, condition SC3, is so strong that no cause that stands in a probabilistic causal relationship to its effect is able to fulfill it. As an illustration, consider figure 3.1, which shows the probabilistic forest fire scenario from section 2.4.4. We assume to know that $A=1$, $L=1$, and $F=1$.

We already showed that, after updating on $F=1$, we end up with a probabilistic SEM $(\mathcal{M}^F,\mathcal{P})$, in which $\mathcal{P}$ assigns a probability of 0 to all contexts $\vec{u}=(u_1,u_2,u_{AF},u_{LF})$, except for:

- $\mathcal{P}(\vec{u_0}=(1,1,1,1))=0.538$
- $\mathcal{P}(\vec{u_1}=(1,1,1,0))=0.231$

[2] The acronym 'PROSAC' simply stands for 'probability of being a strong actual cause'.

- $\mathcal{P}(\vec{u_2} = (1,1,0,1)) = 0.231$



- $A := U_1$
- $L := U_2$
- $F := (A \wedge U_{A,F}) \vee (L \wedge U_{L,F})$

Figure 3.1: $\mathcal{M}^F$ - the probabilistic forest fire scenario.

Intuitively, we have two potential explanations of $F = 1$ in this scenario: We know that $A = 1$ could have causally produced $F = 1$ and we know that $L = 1$ could have causally produced $F = 1$. But neither $A = 1$ nor $L = 1$ is (part of) a sufficient cause of $F = 1$ in any of the three contexts, or even in any logically possible context for $\mathcal{M}$ at all. This is due to condition SC3, which demands that the effect of the cause candidate must hold in every logically possible causal setting for $\mathcal{M}$, in which the cause candidate is brought about by an intervention. But take, for example, the logically possible context $\vec{u_4} = (0,0,0,0)$. Since $U_{A,F} = 0$ and $U_{L,F} = 0$ hold in that context, we can set A to 1, L to 1, or even both, without obtaining $F = 1$. Therefore, neither $A = 1$, $L = 1$, nor $A = 1 \wedge L = 1$ satisfy SC3.

This is clearly a general problem. Imagine that the causal relationship between a cause candidate $X = x$ and an effect $Y = y$ is probabilistic, which means that the causal model representing this relationship contains an error-term $U_{X,Y}$ that represents whether $X = x$ would produce $Y = y$, given that $X = x$ is the case. Then $X = x$ cannot be a sufficient cause of $Y = y$ in any setting for the causal model, because there is always a logically possible context, in which the error-term $U_{X,Y}$ takes on the value 0, while no other cause is present to ensure that $Y = y$ is the case. Sufficient causation, as defined in chapter 1, is therefore incompatible with probabilistic causal relationships.

We can now corroborate the claim from section 1.5.7, namely that the concept of sufficient causation is a bad choice for defining concepts of causal explanation. If we would base our amended definitions of explanation on the concept of sufficient causation, we would be faced with the following dilemma: Either, we would have to accept that an event $\vec{X} = \vec{x}$ that stands in a probabilistic causal relationship to another event ϕ can never explain ϕ, or we would have to embrace the position that there are two distinct types of explanation: One type of explanation for deterministic causal relationships, that is based on the concept of sufficient causation, and another type of explanation for probabilistic causal relationships. I consider both options to be highly unwarranted.

The concept of strong actual causation, on the other hand, is compatible with probabilistic causal relationships. The effect of a strong actual cause does not have to hold in every logically possible setting for the given causal model, in which the strong actual cause is realized by an intervention. Instead, for $\vec{X} = \vec{x}$ to be a strong actual cause of ϕ in a given causal setting $(\mathcal{M}, \vec{u})$, $\vec{X} = \vec{x}$ only has to be sufficient for ϕ in $(\mathcal{M}, \vec{u})$ itself, in the sense that ϕ continues to hold, even if we arbitrarily change the values of all endogenous variables that do not lie on an active causal path from the variables in $\vec{X}$ to the variables in ϕ. This sufficiency does not have to hold in any other setting for $\mathcal{M}$ besides $(\mathcal{M}, \vec{u})$. This is why a cause that stands in a probabilistic causal relationship to its effect can very well satisfy all conditions of strong actual causation. Take, for example, $A = 1$ in $(\mathcal{M}^F, \vec{u_1})$. $A = 1$ is an actual cause of $F = 1$ in $(\mathcal{M}^F, \vec{u_1})$. It thereby already fulfills conditions SAC1 and SAC2. Now, L is the only endogenous variable not lying on an active causal path from A to F. But, whether we set L to 1 or to 0, $F = 1$ continues to hold as long as we keep A on 1. So, $A = 1$ is indeed a strong actual cause of $F = 1$. By analogous reasoning, we get that $L = 1$ is a strong actual cause of $F = 1$ in $(\mathcal{M}^F, \vec{u_2})$. And in $(\mathcal{M}^F, \vec{u_0})$ both $A = 1$ and $L = 1$ are strong actual causes of $F = 1$. We can therefore also determine the respective probabilities of being a strong actual cause in the given probabilistic SEM $(\mathcal{M}^F, \mathcal{P})$. We have $\mathcal{P}(A = 1 \rightarrowtail F = 1) = \mathcal{P}(\vec{u_0}) + \mathcal{P}(\vec{u_1}) = 0.769$, and $\mathcal{P}(L = 1 \rightarrowtail F = 1) = \mathcal{P}(\vec{u_0}) + \mathcal{P}(\vec{u_2}) = 0.769$.

3.4 Explanations in Probabilistic Causal Scenarios

After having established that and how the concept of strong actual causation can be employed in probabilistic SEMs, we can now take on the main task of this chapter: The explication of causal explanation in probabilistic SEMs. A look at our definitions in chapter 1 quickly reveals that there is not much that we have to revise. The concepts of explanation that we defined in chapter 1 are all considered to be relative to an epistemic state of a certain agent: Any explanation is an explanation for someone. This is why our criteria for being an explanation do not simply refer to a causal setting $(\mathcal{M}, \vec{u})$, which describes a real-life causal scenario, but to an agent's epistemic state concerning the causal scenario. As illustrated in chapter 1, this epistemic state concerning a causal scenario can be represented by a causal model $\mathcal{M}$ and a probability distribution $\mathcal{P}$ over all logically possible contexts for $\mathcal{M}$, where $\mathcal{P}$ encodes the agent's degrees of beliefs about which context is the actual one. But this combination of a causal model $\mathcal{M}$ and a probability distribution $\mathcal{P}$ over all logically possible contexts for $\mathcal{M}$ is nothing else than a probabilistic SEM. From a formal perspective, our amended definitions of explanation as formulated in chapter 1 are therefore already accomodated in the framework of probabilistic SEMs.

Notice, though, that our concepts of explanations cannot be defined relative to any kind of probabilistic SEM. As discussed in chapter 2, a probabilistic SEM $(\mathcal{M}, \mathcal{P})$ allows for several interpretations of the probability distribution $\mathcal{P}$. Even if $(\mathcal{M}, \mathcal{P})$ is understood to represent a token scenario, $\mathcal{P}$ can either be interpreted epistemically, as representing degrees of beliefs, or as an objective chance function. Our conceptions of explanation, on the other hand, are considered to be epistemically relative. Any explanation is understood to be relative to a given epistemic state. Therefore, when we define our concepts of explanation in the framework of a probabilistic SEM $(\mathcal{M}, \mathcal{P})$, the probability distribution $\mathcal{P}$ should always be interpreted epistemically, that is as representing the degrees of beliefs of a certain agent α.

In chapter 1, we often represented an agent's epistemic state by a set $\mathcal{K}$ of contexts $\vec{u}_i$ for a causal model $\mathcal{M}$. As already pointed out, a probability distribution $\mathcal{P}$ over the contexts for $\mathcal{M}$ is just a more detailed description of an agent's epistemic state, since it describes the uncertainties of this agent not only qualitatively, but quantitatively. Any epistemically interpreted distribution $\mathcal{P}$ induces a qualitative description $\mathcal{K}^{\mathcal{P}}$ of this very same epistemic state in the following way: $\mathcal{K}^{\mathcal{P}} = \{\vec{u}_i : \mathcal{P}(\vec{u}_i) > 0\}$. We can now restate our explanation definitions in the following way:

Potential Explanation. *$\vec{X} = \vec{x}$ is a 'potential explanation' of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$ if and only if the following conditions hold:*

E1 There is an event $\vec{S} = \vec{s}$ with $\vec{X} = \vec{x}$ being a part of $\vec{S} = \vec{s}$ and $\mathcal{P}(\vec{S} = \vec{s} \rightarrowtail^{\mathcal{M}} \phi) > 0$.

E2 For all contexts $\vec{u} \in \mathcal{K}^{\mathcal{P}}$: $(\mathcal{M}, \vec{u}) \models \vec{S}' = \vec{s}'$ (where $\vec{S}' = \vec{S} \setminus \vec{X}$ and $\vec{s}'$ is the restriction of $\vec{s}$ to the variables in $\vec{S}'$).

E3 $\mathcal{P}(\vec{S} = \vec{s} \rightarrowtail^{\mathcal{M}} \phi) < 1$.

Actual Explanation. *$\vec{X} = \vec{x}$ is an 'actual explanation' of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$ if and only if it satisfies E1, E2, and:*

E4 $\mathcal{P}(\vec{S} = \vec{s} \rightarrowtail^{\mathcal{M}} \phi) = 1$.

Parsimonious Potential Explanation. *A potential explanation $\vec{X} = \vec{x}$ of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$ is 'parsimonious' if and only if it satisfies E1, E2, E3, and:*

E5 $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies E1 and E2.

Explanation. *We will simply say that $\vec{X} = \vec{x}$ is an 'explanation' of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$ if and only if $\vec{X} = \vec{x}$ is either a potential or an actual explanation of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$, that is, if it satisfies E1 and E2 relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$.*

Here again, we will say that an explanation of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$ is 'explicitly complete' if and only if it satisfies E1, E2 and $\vec{S}' = \varnothing$, that is: $\vec{X} = \vec{x}$ and $\vec{S} = \vec{s}$ are identical. With our new terminology, this can be simplified to:

Explicitly Complete Explanation. *$\vec{X} = \vec{x}$ is an 'explicitly complete explanation' of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$ if and only if $\mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} \phi) > 0$.*[3]

In addition to complete and explicitly complete explanations, we can again define the concepts of partial and ambivalent partial explanations:

Partial Explanation. *Let $\mathcal{K}^{\mathcal{P}}_{\vec{X}=\vec{x},\phi}$ be the largest subset of $\mathcal{K}^{\mathcal{P}}$ such that $\vec{X} = \vec{x}$ is an explanation of ϕ relative to $\mathcal{K}^{\mathcal{P}}_{\vec{X}=\vec{x},\phi}$ in $\mathcal{M}$. $\vec{X} = \vec{x}$ is a partial explanation of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$ if and only if $\varnothing \neq \mathcal{K}^{\mathcal{P}}_{\vec{X}=\vec{x},\phi} \subset \mathcal{K}^{\mathcal{P}}$. The degree of explanatory completeness of the partial explanation $\vec{X} = \vec{x}$ is given by:*

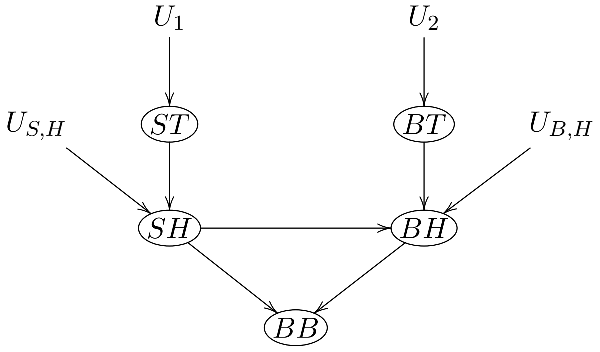
[3] If and only if $\mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} \phi) = 1$, then $\vec{X} = \vec{x}$ is an 'explicitly complete actual explanation' of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$. If and only $0 < \mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} \phi) < 1$, then $\vec{X} = \vec{x}$ is an 'explicitly complete potential explanation' of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$.

$$\mathcal{P}(\mathcal{K}^{\mathcal{P}}_{\vec{X}=\vec{x},\phi}) = \sum_{\vec{u}_i \in \mathcal{K}^{\mathcal{P}}_{\vec{X}=\vec{x},\phi}} \mathcal{P}(\vec{u}_i) \tag{3.2}$$

Ambivalent Partial Explanation. *If and only if there are at least two largest, non-empty, real subsets $\mathcal{K}^{\mathcal{P}}_{\vec{X}=\vec{x},\phi}$, $\mathcal{L}^{\mathcal{P}}_{\vec{X}=\vec{x},\phi}$ of $\mathcal{K}^{\mathcal{P}}$ such that $\vec{X} = \vec{x}$ is an explanation of ϕ relative to $\mathcal{K}^{\mathcal{P}}_{\vec{X}=\vec{x},\phi}$ and relative to $\mathcal{L}^{\mathcal{P}}_{\vec{X}=\vec{x},\phi}$ in $\mathcal{M}$, then $\vec{X} = \vec{x}$ is an ambivalent partial explanation of ϕ relative to $\mathcal{K}^{\mathcal{P}}$ in $\mathcal{M}$.*

It is easy to see that all these definitions are equivalent to the definitions we have given in chapter 1. So, our definitions of explanation remain exactly as stated in chapter 1 and can be employed in contexts with deterministic causal relationships just as well as in contexts with probabilistic causal relationships.

As an illustration consider a probabilistic version of the bottle breaking scenario, in which we assume the relation between Suzy's and Billy's throws and their respective hits to be probabilistic. The scenario can be represented by the SEM $\mathcal{M}^B$ as shown in figure 3.2.



- $ST := U_1$
- $BT := U_2$

- $SH := ST \wedge U_{S,H}$
- $BH := (BT \wedge U_{B,H}) \wedge \neg SH$
- $BB := SH \vee BH$

Figure 3.2: $\mathcal{M}^B$ - the probabilistic bottle breaking scenario.

Let us assume that we already know that $ST = 1$, $BT = 1$, and $BB = 1$. But we do not know who has hit the bottle. So, we have: $\mathcal{P}_{pre}(U_1 = 1) = \mathcal{P}_{pre}(U_2 = 1) = 1$. Let us also assume that we have assembled statistical data on Suzy's and Billy's marksmanship that indicate the following values: $\mathcal{P}_{pre}(U_{S,H} = 1) = 0.8$, $\mathcal{P}_{pre}(U_{B,H} = 1) = 0.2$. With $\vec{u} = \{u_1, u_2, u_{SH}, u_{BH}\}$, this gives us the following preliminary distribution:[4]

- $\mathcal{P}_{pre}(\vec{u}_0 = (1,1,0,1)) = 0.2 \times 0.2 = 0.04$
- $\mathcal{P}_{pre}(\vec{u}_1 = (1,1,1,0)) = 0.8 \times 0.8 = 0.64$
- $\mathcal{P}_{pre}(\vec{u}_2 = (1,1,1,1)) = 0.8 \times 0.2 = 0.16$

[4] Here again, we only need the probabilities for those contexts $\vec{u}_i$ with $(\mathcal{M}, \vec{u}_i) \models ST = 1 \wedge BT = 1 \wedge BB = 1$.

We can now determine the updated probability distribution $\mathcal{P}(\cdot) = \mathcal{P}_{Pre}(\cdot \mid BB = 1)$ for each of these contexts:

- $\mathcal{P}(\vec{u_0}) = \frac{\mathcal{P}_{pre}(\vec{u_0})}{\sum_{i=0}^{2} \mathcal{P}_{pre}(\vec{u_i})} = \frac{0.04}{0.84} \approx 0.05$
- $\mathcal{P}(\vec{u_1}) = \frac{\mathcal{P}_{pre}(\vec{u_1})}{\sum_{i=0}^{2} \mathcal{P}_{pre}(\vec{u_i})} = \frac{0.64}{0.84} \approx 0.76$
- $\mathcal{P}(\vec{u_2}) = \frac{\mathcal{P}_{pre}(\vec{u_1})}{\sum_{i=0}^{2} \mathcal{P}_{pre}(\vec{u_i})} = \frac{0.16}{0.84} \approx 0.19$

Now, it is easy to check that $ST = 1$ is a strong actual cause of $BB = 1$ in $\vec{u}_1$ and in $\vec{u}_2$. We therefore have $\mathcal{P}(ST = 1 \rightarrowtail^{\mathcal{M}} BB = 1) = \mathcal{P}(\vec{u}_1) + \mathcal{P}(\vec{u}_2) \approx 0.95$. Also, since $BT = 1$ is a strong actual cause of $BB = 1$ in $\vec{u}_0$, we have $\mathcal{P}(BT = 1 \rightarrowtail^{\mathcal{M}} BB = 1) = \mathcal{P}(\vec{u}_0) \approx 0.05$. This means that $ST = 1$ as well as $BT = 1$ are potential explanations of $BB = 1$. Both explanations are parsimonious and explicitly complete. But clearly, the probability that $ST = 1$ amounts to an actual explanation of $BB = 1$ (0.95) is much higher than the probability that $BT = 1$ amounts to an actual explanation of $BB = 1$ (0.05). This is due to the fact that Suzy has a much higher marksmanship than Billy and that Suzy's throws harder than Billy and therefore preempts Billy's stone from hitting the bottle, if she herself hits the bottle.

Now imagine that we now come to learn that Suzy did not hit the bottle. We can then update our distribution by conditioning on $SH = 0$. Doing so leaves us with the following distribution $\mathcal{P}'(\cdot) = \mathcal{P}(\cdot|SH = 0)$:

- $\mathcal{P}'(\vec{u}_0 = (1, 1, 0, 1)) = 1$,

such that $\mathcal{P}'$ ascribes a probability of 0 to all contexts besides $\vec{u}_0$. This means that $\mathcal{P}'(ST = 1 \rightarrowtail^{\mathcal{M}} BB = 1) = 0$ such that $ST = 1$ is no potential explanation of $BB = 1$ anymore. But we now have $\mathcal{P}'(BT = 1 \rightarrowtail^{\mathcal{M}} BB = 1) = 1$ such that $BT = 1$ is an explicitly complete actual explanation of $BB = 1$. This is a reasonable result, since we now know that only Billy's throw could have caused the bottle to break.

3.5 Summary

We are now equipped with an account of causal explanation that is applicable in contexts with deterministic causal relationships as well as in contexts with probabilistic causal relationships. For now, this concludes the exposition of my new account of causal explanation.

In the next chapter, which introduces the second part of my dissertation, I will examine two concepts that have already played an important role in the preceding discussions, but whose precise meaning I have not yet discussed: The concept of causal production and the concept of the strength or power of a probabilistic causal relationship. A more thorough investigation of these concepts will not only lead us to an explication of explanatory power, it will also help us to better understand why the concept of strong actual causation is such a fitting candidate for defining causal explanation.

Part II

Causal Explanatory Power

Chapter 4

Causal Power

4.1 Introduction

The structural equations in an SEM and the probabilities of error-terms encode crucial information about the existence and the strength of causal relationships. So far, we have simply taken this information as given. But how exactly do we discover causal relationships and how can we evaluate their strength?

In the present chapter, I will explore a theory of causal induction that aims to answer these questions. The theory has been developed by the psychologist Patricia Cheng. While there are other elaborate accounts of causal induction, especially those developed by Pearl (2000) and Spirtes et al. (2000) in the causal model framework, I focus on Cheng's theory, because it yields a measure of causal power that will ultimately prove to be very useful for the project of this dissertation.

Recently, though, Cheng's measure of causal power has been confronted with severe challenges. Sprenger (2018) has put forward several arguments, according to which Cheng's measure is completely inadequate as a measure of causal strength. Although Sprenger's arguments are based on formally sound results, I will argue that these results do not actually suggest to dismiss Cheng's measure. Instead, they should serve as a motivation to generalize Cheng's theory of causal induction and the measure of causal power that results from it. By exploring and expanding Cheng's theory, we will also make some crucial advances in grasping a notion that has permeated our discussions about causation from the very beginning of this dissertation, namely the concept of causal production.

Before starting the chapter properly, a notification concerning the notation is due. In the previous chapters, I have used upper case roman letters to denote variables in a causal model. But in the present chapter, we do not work within the causal model framework. I will therefore use upper case roman letters, like 'C' and 'E', to denote event-types or token events. And I will use 'P' to denote a probability function.

4.2 Cheng's Power PC Theory

It is already a platitude, though an important one, that correlation cannot simply be identified with causation. We can neither infer from a correlation between two event types C and E

that there is a direct causal relation between C and E nor can we infer that there will be a correlation between two event types C and E when there is a causal relationship between them. Still, it has long been acknowledged that there is some deep connection between correlation and causation. Reichenbach's principle of common cause can serve as a famous example. Closely linked with this assumption is the well entrenched idea that, at least under certain, well curated, circumstances we can use information about the correlation of event-types to infer something about their causal relationship. This conviction also underlies Patricia Cheng's Power PC theory, which, as a psychological theory of causal induction, aims to describe how human beings actually induce the presence and the strength of an unobservable causal relationship between event-types C and E from the observable correlation or covariation of C and E. The nucleus of Cheng's theory is nicely summarized in the following passage:

> "the reasoner believes that there are such things in the world as causes that have the power to produce an effect and causes that have the power to prevent an effect and that only such things influence the occurrence of an effect" (Cheng, 1997, p. 372).

Even though causation itself is understood to be unobservable, the Power PC theory assumes that human beings employ certain causal concepts as theoretical terms and impose them on the observable realm to explain observable phenomena like the occurrences of events and the covariation of event-types. According to Cheng's theory, the following four causal concepts are particularly important: Causal production, generative causal power ("the power to produce an effect"), causal prevention, and preventive power ("the power to prevent an effect").[1] In the present chapter, I will focus on causal production and generative causal power. Causal prevention and preventive power will be the topic of the next chapter. In Cheng's theory, the observable covariation between event-types C and E is always assumed to be given by the following term:

$$\Delta P_{C,E} = P(E|C) - P(E|\neg C) \tag{4.1}$$

where the probabilities are understood to be relative frequencies that are ideally obtained from some randomized controlled trial (RCT).[2] Now, as human beings, we have the tendency to explain an observable covariation $\Delta P_{C,E}$ by some kind of causal relationship, for example, by assuming that events of type C have a certain amount of generative causal power $g_{C,E}$ on events of type E. But the crucial questions are: Under what conditions are we allowed to infer the existence of a causal power from an observed covariation? And how exactly is the quantity $g_{C,E}$ of causal power related to the magnitude $\Delta P_{C,E}$ of covariation? According to the Power PC theory, the relation between $g_{C,E}$ and $\Delta P_{C,E}$ is ultimately determined by several a priori assumptions about the theoretical causal concepts of causal production and generative causal power. These a priori assumptions basically serve as bridge principles between the two unobservable causal concepts, on the one hand, and observable concepts, like the occurrences of token events and

[1] The theoretical causal concepts that feature in Cheng's Power PC theory may very well not denote any ontic objects in the real world. The Power PC theory is no ontological theory about causation. It is a psychological theory of how human beings describe the world and what kind of causal concepts we employ to do so. Still, according to Cheng's theory, the causal concepts are supposed to be intersubjective, since we all make the same a priori assumptions about them. Cheng seems to stay agnostic about the metaphysical question of whether there really are human-independent causal powers in the ontic world. I am going to do the same.

[2] We will later consider some illustrative examples.

the covariation of event-types, on the other hand. It is due to these assumptions that we are, at least in some situations, able to deduce the unobservable causal power $g_{C,E}$ from the observable covariation $\Delta P_{C,E}$. We can think of the a priori assumptions about the theoretical causal concepts as the axioms of the Power PC theory, which give the otherwise empty notions of causal power and causal production their meaning.

The following list of axioms contains some axioms that Cheng (1997) herself does not make explicit. But they are clearly presupposed by her Power PC theory. The following list is similar, though still more comprehensive, to a list of assumptions that Hiddleston (2005, p. 39) puts forward as a description of Cheng's theory. Also, unlike Cheng or Hiddleston, I will subdivide the assumptions into two groups. The first group, C1-C5, is a list of axioms that are supposed to hold rather generally. The assumptions CP1 to CP3, on the other hand, should rather be seen as auxiliary assumptions, that are made to apply the Power PC theory to a very simple kind of causal scenario.

Now, the first axiom is a characterization of causal power:[3]

C1 The generative causal power $g_{C,E}$ of C to produce E can be represented by the probability that C produces E, given that C is the case ($P(C$ produces $E|C)$).

C1 already points to a mathematical representation of generative causal power. But, again, notice that the concept of causal production, which is employed in this characterization of causal power, is a theoretical, unobservable term, which gets its meaning only from the following assumptions, which connect events of causal production with observable occurrences of events:[4]

C2 If C produces E, C is the case and E is the case.

The first part of C2, namely that C can only produce E, if C is the case, is quite intuitive. The second part aims to ensure that causal production is a success term: It is impossible that C produces E without E actually happening. We further have:[5]

C3 Any occurring event E is causally produced.

The next axiom is:[6]

C4 Any event E can be produced by arbitrarily many events, i.e. causal overdetermination is possible for any E.

And finally:

[3] See (Cheng, 1997, p. 372). For now, we presuppose that there is no prevention involved. When we have introduced the concept of prevention, we will see that C1 is in need of adjustment.

[4] Cheng (1997) does not make assumption C2 explicit, but Hiddleston (2005, p. 39) does.

[5] Humphreys (1989) points out that such an assumption might be too strong. There could actually be events, like the decay of a nucleus, that could happen without any cause. Events like these simply have a base rate probability of being instantiated. We could incorporate this idea into the Power PC theory by treating C3 as an axiom that is not supposed to hold generally, but only for a certain kind of events. But in the following, I will assume that C3 holds generally.

[6] Cheng does not state C4 explicitly, but it follows from CP6, which is stated explicitly by Cheng (1997). I state C4 as an axiom of its own, because CP6 will not be assumed to hold generally in the generalized Power PC theory that I will later propose.

C5 For every generative cause C of E, the magnitude of the generative causal power $g_{C,E}$ is independent from the probability of C.

We now come to the second group of assumptions, which, as pointed out above, should not be understood to hold generally, since they describe a specific, though very simple, kind of causal scenario, to which the Power PC theory is then applied. The first assumption ensures that we are dealing with a scenario with an event E for which there are only two factors that can causally influence it: the event C with the generative causal power $g_{C,E}$ on E and some known or unknown, potentially composite event U with the generative causal power $g_{U,E}$ on E:[7]

CP1 If C is the case, it produces E with probability $g_{C,E}$ and if U is the case, it produces E with probability $g_{U,E}$ and nothing else can influence the occurrence of E.

Cheng (1997, p. 373) further assumes:

CP2 $g_{C,E}$ is independent from the probability of U and from $g_{U,E}$.

CP3 U and C are probabilistically independent.

The kind of scenario described by CP1-CP3, which I will denote a *Cheng-scenario*, can be represented by a graph as shown in figure 4.1.

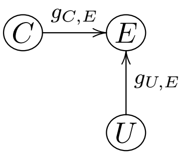


Figure 4.1: A Cheng-scenario.

Now, the assumptions, that we have made, ensure that in a Cheng-scenario the probability of E is given by:[8]

$$P(E) = P(C) \times g_{C,E} + P(U) \times g_{U,E} - P(C) \times g_{C,E} \times P(U) \times g_{U,E} \tag{4.2}$$

This equation already entails the mathematical relation between the causal power $g_{C,E}$ and the observable covariation $\Delta P_{C,E} = P(E|C) - P(E|\neg C)$ of C and E. To obtain $\Delta P_{C,E}$, we first have to conditionalize equation (4.2) on C. Since $P(C|C) = 1$ and since $g_{C,E}$, $g_{U,E}$, and $P(U)$ are independent of $P(C)$, we obtain:

$$P(E|C) = g_{C,E} + P(U) \times g_{U,E} - g_{C,E} \times P(U) \times g_{U,E} \tag{4.3}$$

Next, we conditionalize equation (4.2) on $\neg C$. Since $P(C|\neg C) = 0$, we obtain:

$$P(E|\neg C) = P(U) \times g_{U,E} \tag{4.4}$$

With $\Delta P_{C,E} = P(E|C) - P(E|\neg C)$ we get:

[7] See (Cheng, 1997, p. 373).

[8] See (Cheng, 1997, p. 373-4) for the following deduction of Cheng's measure of generative causal power.

$$\Delta P_{C,E} = g_{C,E} - g_{C,E} \times P(U) \times g_{U,E} = g_{C,E} \times (1 - P(U) \times g_{U,E}) \tag{4.5}$$

Solving equation (4.5) for $g_{C,E}$ gives us:

$$g_{C,E} = \frac{\Delta P_{C,E}}{1 - P(U) \times g_{U,E}} \tag{4.6}$$

Equation (4.4) tells us that $P(E|\neg C) = P(U) \times g_{U,E}$, which gives us:

$$g_{C,E} = \frac{\Delta P_{C,E}}{1 - P(E|\neg C)} = \frac{P(E|C) - P(E|\neg C)}{1 - P(E|\neg C)} \tag{4.7}$$

Equation (4.7) reaches the main target of Cheng's Power PC theory: It describes how the unobservable causal power $g_{C,E}$ of C on E can be determined by observable relative frequencies, namely $P(E|C)$ and $P(E|\neg C)$ – at least, as long as we are dealing with a Cheng-scenario. The formula on the right-hand side of equation (4.7) is what has come to be known as *Cheng's measure of generative causal power*, which, in the following, I will denote by 'CS_{ch}'.

Now, one crucial question, that Cheng has left untouched, remains: What exactly is the generative causal power of C on E? Equation (4.7) equates $g_{C,E}$ with a ratio of relative frequencies. But this does not necessarily mean that $g_{C,E}$ *is* a ratio of relative frequencies. It only means that its value can be determined by a ratio of relative frequencies. Similarly, axiom C1 equates $g_{C,E}$ with a certain conditional probability. But here again, this does not mean that $g_{C,E}$ *is* a conditional probability. It only means that the value of the conditional probability $P(C \text{ produces } E|C)$ is equal to the value of $g_{C,E}$.[9] I want to make the following proposal: Even though the value of $g_{C,E}$ can be determined by certain probabilities, $g_{C,E}$ is not a probability itself. It is a power, a capacity of an event of a certain type to produce tokens of another event-type. This capacity comes in degrees that, just like probabilities, can be measured within the scale $[0, 1]$. But even though the causal power $g_{C,E}$ of an event C to produce another event E is no probability itself, it still has an intricate connection to probabilities in all its different interpretations. Causal powers can be seen as the reason why certain probabilities emerge.[10] For example, the causal power $g_{C,E}$ of tokens of event-type C can explain why certain relative frequencies emerge. Imagine that we have a certain number k of C-tokens, each with a causal power $g_{C,E} = x$ to produce an event of type E and imagine that no other potential causes of E are present. We then have k C-tokens, each with a probability of x to causally produce E. According to the weak law of large numbers, as k approaches infinity, the relative frequency $rel(E|C)$ converges on x with probability 1. Or, put differently, the larger k, the higher the probability of $rel(E|C)$ being (close to) x.[11] But while the causal power $g_{C,E}$ can give rise to

[9] We will later see that the conditional probability $P(C \text{ produces } E|C)$ does not always determine the value of $g_{C,E}$. It only does so under certain conditions that we will explore in more detail later.

[10] Our understanding of $g_{C,E}$ is therefore reminiscent of certain explications of *propensity*. See (Berkovitz, 2015) for an overview of different explications of this concept.

[11] This illustrates an important idealization of Cheng's measure of generative causal power. Observable, that is finite, relative frequencies can never determine a causal power with absolute reliability. We can only reach absolute accuracy and reliability when we use "ideal", that is infinite, relative frequencies, where k reaches infinity. As long as k is finite, the resulting relative frequencies $rel(E|C)$ and $rel(E|\neg C)$ can only determine the value of the causal power $g_{C,E}$ approximately. For the rest of this dissertation I will adopt the idealization made by Cheng's measure and I will simply assume that all observable relative frequencies are actually identical to the respective ideal frequencies.

certain relative frequencies, it is no relative frequency itself. A crucial difference between causal powers and relative frequencies is that single case token events can have non-trivial values of causal power, but they cannot have non-trivial relative frequencies.

The causal power $g_{C,E}$ of a token event C to produce another event E should also influence the credence of a rational agent α about whether E is the case, given that C is the case. In analogy to the principle of direct probability and the principal principle, we can formulate the following guideline with cr_α being a credence-function relative to agent α: If α knows that there is no other potential cause of E present and α does not have any trumping evidence about whether the token event E happens or not, then:

$$cr_\alpha(E|C \wedge g_{C,E} = x) = x \tag{4.8}$$

Yet again, even though $g_{C,E}$ can give rise to certain credences, it is no credence itself. Credences about events are agent-dependent and always relative to a given epistemic state. They are no intrinsic properties of the events themselves. Causal powers on the other hand are supposed to be exactly that: Mind-independent properties of the events themselves.

4.3 Alternative Measures of Causal Strength

Cheng's measure of causal power is not the only measure of causal strength out there. Actually, there is quite a variety of different proposals. Fitelson and Hitchcock (2011) give a nice overview of different probabilistic measures of causal strength that have been proposed in the literature.[12]

Eells' (1991) Measure:	$CS_e(E,C) = P(E|C) - P(E|\neg C)$
Suppes' (1970) Measure:	$CS_s(E,C) = P(E|C) - P(E)$
Galton Measure:	$CS_g(E,C) = 4 \times P(C) \times P(\neg C) \times (P(E|C) - P(E|\neg C))$
Cheng's (1997) Measure:	$CS_{ch}(E,C) = \dfrac{P(E|C) - P(E|\neg C)}{1 - P(E|\neg C)}$
Lewis' (1986a) Measure:	$CS_l(E,C) = \dfrac{P(E|C) - P(E|\neg C)}{P(E|C)}$

Table 4.1: Popular causal strength measures. Adapted from (Fitelson and Hitchcock, 2011).

Notice that the way the measures are presented in table 4.1 is a slight simplification. It is a crucial presupposition for all these measures that confounding is precluded. To achieve this, it is commonly presupposed that when $P(E|C)$ and $P(E|\neg C)$ are determined, the 'background context' remains constant. The background context essentially comprises all other factors that causally influence E, besides those events that lie on an active causal path from C (or $\neg C$) to E. A more accurate description for $CS_e(E,C)$, for example, would therefore be this: $CS_e(E,C) =$

[12] The original list by Fitelson and Hitchcock (2011) additionally contains a measure by Good (1961a, 1961b). I have omitted the measure in table 4.1, because it is ordinally equivalent to Cheng's measure.

$P(E|C \wedge B_i) - P(E|\neg C \wedge B_i)$, with B_i representing the background context. Alternatively, we can think of $P(E|C)$ and $P(E|\neg C)$ not as conditional probabilities, but as the probabilities of E after realizing C, and $\neg C$ respectively, by interventions. In scientific practice, RCTs are a common way of implementing such interventions.[13]

Now, when I call all these different formulas 'measures of causal strength', then this gives the impression that they all measure one and the same thing, namely causal strength. But this is highly misleading. In fact, all these different measures measure different kinds of causal strength. It does therefore not make sense to ask, which of those measures are correct and which are incorrect in measuring *the* strength of a cause on its effect. Just talking about 'the causal strength' of a cause on its effect is ambiguous. Any of the five measures correctly measures its own kind or conception of causal strength. For example, applying Eells' measure to determine the causal strength of a probabilistic causal relationship between C and E may very well yield a different result than applying Suppes' measure. But this is no contradiction, since both measures quantify different concepts or kinds of causal strength, which I will differentiate by different indexes: Eells' measure quantifies the causal strength CS_e between C and E. Suppes' measure, on the other hand, quantifies the causal strength CS_s between C and E.

Nonetheless, just because several different conceptions or kinds of causal strength can peacefully coexist, this does not mean that they are all equally useful. Depending on the stated goals and purposes, certain measures can very well be normatively superior to others, simply by having certain features which are practical for the given purposes and that other conceptions of causal strength lack. Additionally, a certain conception of causal strength may be descriptively superior to other conceptions of causal strength by being more in accord with how human beings actually evaluate the strength of causes. These two dimensions of evaluating and comparing different conceptions of causal strength, the normative and the descriptive dimension, are not completely independent from each other. Given that human beings are pretty successful causal reasoners, any indication that a certain concept of causal strength descriptively matches how human beings with certain goals actually evaluate causal relationships is prima facie evidence for this concept's normative valence for these very same goals.[14]

It would go beyond the scope of this dissertation to compare the value of all five concepts of causal strength for any specific goal a human being might ever have. But we can focus on what seems to be a prototypical or very common use of causal knowledge. And indeed, there is a certain application of causal knowledge, whose gains are so broad and extensive, that it would be fair to view it as the most essential and fundamental purpose for the deployment of causal knowledge: making predictions. As humans, we constantly make predictions based on our causal knowledge and beliefs. Ensuing from given events we make predictions about their potential effects just for knowing what to expect from the future. We also make predictions about the potential effects of alternative, feasable actions to make foresighted and rational decisions. And, as we have seen in chapter 1 and 3, we can also make predictions ensuing from potential causes that may have happened in the past to find potential explanations for a given event.

[13] For Cheng's measure CS_{ch}, we have built the presupposition, that confounding is precluded, into the assumptions about the Cheng-scenario.

[14] Conversely, there may also be certain purposes for which a descriptively very accurate concept of causal strength is especially fruitful or useful.

Making predictions about potential effects of present, hypothetical, or future events is therefore not only essential for foreseeing the future, but also for rational decision making, for choosing an appropriate action for a desired outcome, and for finding potential explanations of given events.

But predictions based on knowledge about causal relationships are only possible, if we assume that certain properties of the causal relationships remain stable or invariant across different contexts and can therefore be projected or extrapolated to new situations. If we want to make a prediction about whether a certain cause C will yield a certain effect E in a certain situation a and we want to base our prediction on the strength $CS(C, E)$ of the causal relationship between C and E, then we cannot measure $CS(C, E)$ in situation a itself. As pointed out above, for determining the strength of a causal relationship between C and E, we need a well-designed context, ideally a randomized controlled trial, in which we observe and count the number of C-tokens, $\neg C$-tokens, E-tokens and $\neg E$-tokens. I will call such a context, in which we are able to measure the causal strength $CS(C, E)$, a learning-context. The typical prediction-context cannot be a learning-context, for the simple reason that we have not yet observed whether the effect E happens. So, whenever we want to base a prediction about the effect E of an event C in a given situation a on the strength $CS(C, E)$ of the causal relationship between C and E, then we have to assume that the value of $CS(C, E)$, as it was determined in the learning-context, remains stable or invariant during contextual changes. We must be able to extrapolate $CS(C, E)$ from the learning-context to the prediction-context.[15] Unless we assume that there is such an invariant causal strength $CS(C, E)$, exact predictions based on causal knowledge are impossible. I will call the type of causal strength $CS(C, E)$, that is assumed to be invariant across different contexts, the *intrinsic causal power* of C on E.[16]

Now, here comes an unsurprising, yet bold and consequential claim: I consider Cheng's concept of causal strength CS_{ch} to be the single best candidate for intrinsic causal power. This claim is unsurprising because I am not the first one to make it. Cheng and her colleagues have expressed this claim in several papers. Additionally, Fitelson and Hitchcock (2011) argue for it. The claim is consequential, because it grants CS_{ch} a very special role among its fellow concepts of causal strength, because without CS_{ch} predictions based on causal knowledge, which is a crucial, if not the crucial application of causal knowledge, would be impossible. The claim is bold, because it is not so easy to back it up. In the following section, I will deal with two arguments for the normative superiority of CS_{ch} as a candidate for intrinsic causal power. The first argument is given by Cheng et al. (2013). But I will argue that it is ill-founded. The second argument is given by Fitelson and Hitchcock (2011) and I will show that it is no conclusive argument either, even though it provides a strong indication that CS_{ch} is a much more promising candidate for intrinsic causal power than other kinds of causal strength. I will ultimately argue that the best argument for the normative superiority of CS_{ch} as an intrinsic measure of causal power is based on its descriptive superiority.

[15] The importance of stable or invariant causal relations for predictions also explains why stable causal relations are typically considered to be much more valuable than instable causal relations. See (Hüttemann, 2021) for a similar argument. Hüttemann values the concept of intrinsic causal power for its capacity to enable extrapolation. Woodward (2006) also argues for the normative superiority of insensitive causal relations. Vasilyeva et al. (2018) show that humans value stable causal relations more than instable causal relations.

[16] I call it *intrinsic*, because it is assumed that extrinsic, contextual factors, that are changes to the context (instead of changes to the cause itself), do not influence or change it.

4.4 CS_{ch} as Intrinsic Causal Power

4.4.1 An Argument for Normative Superiority

Liljeholm and Cheng (2007) argue that CS_{ch} is the only reasonable choice for intrinsic causal power, because treating an alternative kind of causal strength, like CS_e, as intrinsic causal power would lead to a logical contradiction.[17] Before we can go into the details of the argument, we need to introduce a new pair of concepts, namely the concept of *causal interaction* and its antagonist *causal independence*. Liljeholm and Cheng (2007) loosely describe the latter notion like this: "If two or more causes *independently influence* an effect, then when they are jointly present, each operates on the effect as if the other cause or causes were not there" (Liljeholm and Cheng, 2007, p. 1016, *emphasis in original*). This description is still somewhat imprecise, because it does not make explicit, which factor or property of a causal influence should remain stable, whether or not the other cause is present. This is where the intrinsic causal power comes into play. Consider an event C with an intrinsic causal power $CS_{int}(C, E)$ on E and an event D with an intrinsic causal power $CS_{int}(D, E)$ on E. We will say that C causally influences E independently from D iff $CS_{int}(C, E)$ is the same, whether D is present or not. We say that C and D interact in the causal influence on E if either C does not causally influence E independently from D or D does not causally influence E independently from C. So, the meaning of the concepts of causal interaction and causal independence highly depend on the exact meaning of the concept of intrinsic causal power. If we assume, for example, CS_{ch} to be intrinsic, then C causally influences E independently from D if and only if the presence or absence of D does not change the value of $CS_{ch}(C, E)$. But if we assume, CS_e to be intrinsic, then C causally influences E independently from D if and only if the presence or absence of D does not change the value of $CS_e(C, E)$.

Now, Liljeholm and Cheng (2007) claim that the concept of causal independence that would result from assuming CS_e to be intrinsic (I will write "*causal* independence$_e$" for short) is "self-contradictory" (Liljeholm and Cheng, 2007, p. 1016). The same argument is reiterated in (Cheng et al., 2013) and in (Cheng et al., 2007). If this is indeed the case, it would make a very strong case against CS_e being the intrinsic causal power, because, well, self-contradiction should better be avoided. If now the same could be shown not only for CS_e, but also for CS_s, CS_g, and CS_l, we would have a strong argument for CS_{ch} being the single best choice for intrinsic causal power.

Liljeholm and Cheng (2007) use the following example to illustrate their argument. Imagine that we aim to evaluate whether taking a certain medicine A causes headache. Imagine further, that we have two distinct trials, trial a and trial b, to determine the causal strength of taking A on getting a headache. In trial a, we have ensured that there are no potential background causes of getting a headache besides A. In the control-group of trial a (the box on the upper left of Figure 4.2), in which none of the 24 individuals took medicine A, we can therefore observe that none of the individuals developed a headache, which gives us $P(H|\neg A) = 0$.[18] In the test-group

[17] The authors actually only consider CS_e as an alternative to CS_{ch} in their argument. But if the authors' argument would be correct, it could easily be extended to apply to CS_s, CS_l, or CS_g as well.

[18] Notice that I use the variable A here not as standing for the medicine itself, but for the event 'taking medicine A'. H represents the event 'having a headache'.

of trial a (the box on the lower left of Figure 4.1), which also consists of 24 individuals, each individual was administered medicine A and we observe that 6 of the 24 individuals developed a headache. We therefore have $P(H|A) = 0.25$. Trial b is assumed to be done in a completely different context and in this context some background causes of H are present.[19] This is why in the control-group (the box on the upper right of Figure 4.1), in which none of the individuals took medicine A, 16 out of 24 individuals still develop a headache ($P(H|\neg A) = {}^2\!/_3$). In the test-group of trial b (the box on the lower right of Figure 4.1), in which all individuals took medicine A, 22 out of 24 individuals developed a headache ($P(H|A) = {}^{11}\!/_{12}$).

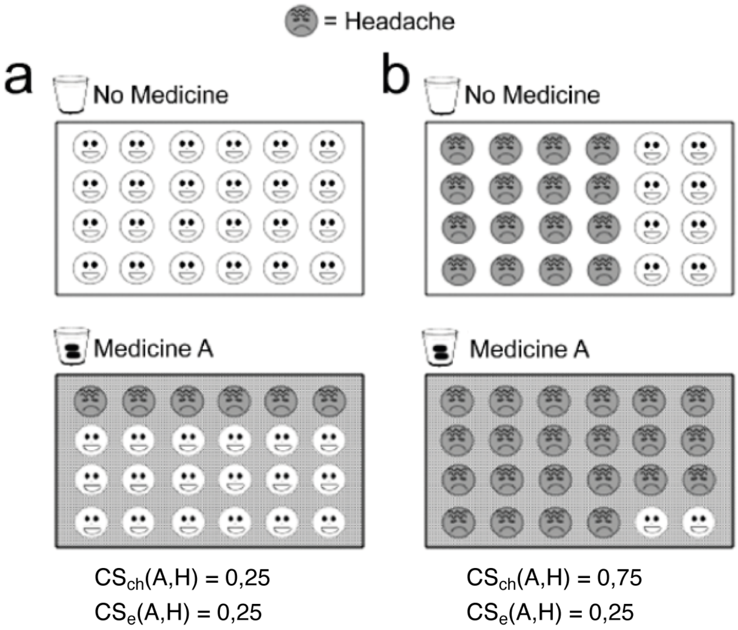


Figure 4.2: RCTs measuring a side-effect of medicine A. Adapted from (Liljeholm and Cheng, 2007, p. 1017).

Now, applying Eells' measure of causal strength, we get $CS_e(A, H) = 0.25$ in trial a as well as in trial b. So, if we assume CS_e to be the intrinsic causal power of A on H, then the stability of $CS_e(A, H)$ across both contexts appears to be evidence that none of the background causes of H, that are present in trial b, interact with A's causal influence on H. According to our definition above, this would mean that A causally influences H independently$_e$ from all background causes of H that are present in trial b.

Liljeholm and Cheng (2007) claim that "for the strength of Medicine A as defined by $[CS_e(A, H)]$ to remain constant across the two studies, those 6 additional patients in the bottom half of Figure [4.2] must not overlap with the 16 who are estimated to have headache due to the background cause; that is, the causes have mutually exclusive influences (i.e., the probability of both causes producing headache in a patient is 0)" (Liljeholm and Cheng, 2007, p. 1016). So, the authors argue that the causal strength $CS_e(A, H)$ of A on H can only remain constant across trial a and trial b, if all the headaches that the 24 tokens of A causally produced in trial b are not already produced by any of the background causes that are present in trial b, that is,

[19] Notice, though, that the prevelance and the strength of these background causes of H are supposed to remain stable across the test-group and the control-group in trial b.

if $P(A$ produces $H \wedge B$ produces $H) = 0$, with B standing for the background causes of H that are present in trial b. But that would mean that A causally influences H independently from B if and only if the causal productions of headaches by A and the causal productions of headaches by B are mutually exclusive. This is where Liljeholm and Cheng (2007) smell logical trouble and they proclaim that: "mutual exclusivity as a definition of independence is self-contradictory" (Liljeholm and Cheng, 2007, p. 1016).[20]

I have two issues with this argument. My first issue concerns the claim that $CS_e(A, H)$ can only be stable across trial a and trial b, if $P(A$ produces $H \wedge B$ produces $H) = 0$. I consider this claim to be wrong. My second issue is that, even if this claim would be true, it would not mean that taking CS_e as the intrinsic causal power would lead to a self-contradicting concept of causal independence.

Let us start with the first issue. The claim that $CS_e(A, H)$ can only be stable across trial a and trial b, if $P(A$ produces $H \wedge B$ produces $H) = 0$ is not self-evident, even though Liljeholm and Cheng (2007) seem to consider it to be so. Instead, we need an additional assumption to warrant its truth, which I will call STP (stability of token-production):

(STP) The causal strength $CS(C, E)$ can only be stable across two contexts, if the same number n of C-tokens causally produce the same number k of E-tokens in both contexts.

In our case, this would mean that $CS_e(A, H)$ can only be stable across trial a and trial b, if the 24 A-tokens in trial a and the 24 A-tokens in trial b both causally produce exactly the same number of H-tokens, namely 6. If we assume this to be true, then $CS_e(A, H)$ can only have the same value in trial b as in trial a, if all the 6 H-tokens, that the 24 A-tokens produce in the test-group of trial b, are not already produced by the background causes. Otherwise the value of $P(H|A)$ would be smaller than $^{11}/_{12}$ and consequently $CS_e(A, H)$ would be smaller than 0.25 in trial b.

STP does indeed hold for CS_{ch}. This is a consequence of how $CS_{ch}(C, E)$ is characterized in Cheng's Power PC theory, namely as being equal to $P(C$ produces $E|C)$. But for the concept of causal strength $CS_e(C, E)$, which is measured by $P(E|C) - P(E|\neg C)$, we have not made any assumptions about its connection to the theoretical term of causal production. This is why it is completely unwarranted to think that STP should hold for CS_e. As soon as we recognize that STP does not have to hold for CS_e, we can see that there is another way for $CS_e(A, H)$ to remain stable across trial a and trial b. $CS_e(A, H)$ can remain stable, even if $P(A$ produces $H \wedge B$ produces $H) = P(A$ produces $H) \times P(B$ produces $H)$, as long as the 24 A-tokens in trial b causally produce more H-tokens than the 24 A-tokens in trial a, namely 18 instead of just 6. One might want to object that it seems very unintuitive to say that a cause A causally influences an effect H independently of other causes B, even though the presence of B has

[20] Notice that we can make similar arguments for the other concepts of causal strength aside from CS_{ch}. The basic (problematic) assumption is always that the respective causal strength can only remain the same across both contexts, if the 24 tokens of A produce just the same amount of headaches in trial b as the 24 tokens of A produce in trial a. But for all concepts of causal strength besides CS_{ch}, this can only be the case if $P(A$ produces $H \wedge B$ produces $H) \neq P(A$ produces $H) \times P(B$ produces $H)$. Only for CS_{ch}, we can have that $P(A$ produces $H \wedge B$ produces $H) = P(A$ produces $H) \times P(B$ produces $H)$, while $CS_{ch}(A, H)$ remains the same across both trials and the 24 tokens of A produce just the same amount of headaches in trial b as the 24 tokens of A produce in trial a.

the consequence, that the same number of A-tokens causally produce more H-tokens than they do in the absence of B. I agree with that. But this only shows that we typically use the concept of causal independence differently and not that the concept of causal independence$_e$ is self-contradictory.

Let us now, just for the sake of the argument, assume that STP holds for CS_e. Does taking CS_e as the intrinsic causal power then automatically lead to a self-contradicting concept of causal independence, as Liljeholm and Cheng (2007) claim? The answer is clearly no. It is no logical contradiction to say that the event 'A produces H' is probabilistically independent from the event 'B produces H', while A causally influences H independently from B. The probabilistic independence of certain production-events is simply a different kind of independence as the independence$_e$ of causal influence. There is no logical contradiction if one of them is present while the other one is absent. Contrary to the claims of Cheng and her colleagues, taking CS_e, or any other concept of causal strength besides CS_{ch}, as intrinsic causal power does not lead to a self-contradicting concept of causal independence. What Cheng and her colleagues consider to be an a priori argument for the superiority of CS_e as intrinsic causal power, is in fact a fallacy that results from a false assumption (STP) and from equating two different things that happen to have the same name.

4.4.2 Another Argument for Normative Superiority

Fitelson and Hitchcock (2011) have pointed out that, except for CS_{ch}, all concepts of causal strength in table 4.1 exhibit, what the authors call, 'floor effects'. This means that the causal strength of C on its effect E in a certain context is restricted by the value of $P(E|\neg C)$ in that context. This is quite easy to see for CS_e, but it equally holds for all other concepts of causal strength in table 4.1, except for CS_{ch}. Imagine, for example, we have an event of type C and an event of type E and we have determined the causal strength $CS_e(C,E)^a$ of C on E in a context a, in which no other potential causes of E are present, such that $P^a(E|\neg C) = 0$. Since $P^a(E|C)$ can in principle take on any value between 0 and 1, the causal strength $CS_e(C,E)^a = P^a(E|C) - P^a(E|\neg C)$ of C on E in context a can in principle have any value between 0 and 1. Now imagine a context b, in which other potential causes of E besides C are present, such that $P^b(E|\neg C) = x > 0$. Since the highest possible value for $P^b(E|C)$ is still 1, the causal strength $CS_e(C,E)^b = P^b(E|C) - P^b(E|\neg C) = P^b(E|C) - x$ of C on E in context b can only take on a value between 0 and $1 - x$. The higher the value of $P(E|\neg C)$ in a given context, the smaller becomes the range of values that the causal strength CS_e of C on E can take on in that context. Put differently: The more potential causes of E, besides C itself, are present in a given context and the stronger those potential causes are, the smaller becomes the range of values that the causal strength CS_e of C on E can take on in that context. Concepts of causal strength that exhibit floor effects, like CS_e, CS_l, CS_s, and CS_g, are therefore, due to their mathematical structure alone, highly context-sensitive.

Imagine, for example, that we determine the causal strength CS_e of C on E in a context a, in which no other potential causes of E are present, with the result that $CS_e(C,E)^a = 0.6$. Now imagine a context b with other potential causes of E besides C such that $P^b(E|\neg C) = 0.5$. We now know a priori that the value of $CS_e(C,E)^b$ cannot be 0.6 anymore. The highest possible value for $CS_e(C,E)^b$ is $1-0.5 = 0.5$, which means that the causal strength CS_e of C on E cannot be the same in both contexts a and b. This illustrates that the potential for being invariant over different contexts is a priori restricted for concepts of causal strength that exhibit floor effects.

Fitelson and Hitchcock (2011) further point out that the values of CS_s and CS_g are additionally restricted by the probability of C itself in a given context. The higher the probability of C in a given context, the smaller becomes the range of values that the causal strength CS_e or the causal strength CS_g of C on E can take on in that context. Just like concepts of causal strength that are sensitive to the value of $P(E|\neg C)$, concepts of causal strength that are sensitive to $P(C)$ have an a priori restricted potential to be invariant across different contexts.

There is no such a priori restriction of the potential invariance of CS_{ch} across different contexts. To see this, remember that the causal strength CS_{ch} of C on E is measured by the following formula:

$$CS_{ch}(E,C) = \frac{P(E|C) - P(E|\neg C)}{1 - P(E|\neg C)} \tag{4.9}$$

As can be seen in this equation, the value of $P(E|\neg C)$ does not restrict the range of possible values for $CS_{ch}(E,C)$ in any way. $CS_{ch}(E,C)$ can take on any value between 0 and 1, no matter how high the value of $P(E|\neg C)$ is.[21]

Now, the question is, what should we deduce from these insights? We have seen that, due to the mathematical structures of the measures of causal strength, CS_{ch} has a higher potential to be invariant across different contexts than the other concepts of causal strength. But this does not conclusively prove that CS_{ch} really is more invariant than the other kinds of causal strength in our actual world. Just because there are no a priori reasons for CS_{ch} to change in different contexts does not mean that it actually does not change in different contexts. The causal structure of our actual world may be such that CS_{ch} ends up being more sensitive to contextual changes than the other types of causal strength, even though there are no a priori reasons for CS_{ch} to change in different contexts. Whether CS_{ch} is more invariant across different contexts in our actual world is ultimately an empirical question that cannot be answered a priori. And it is an empirical question for which a conclusive answer is very hard to get. There are countlessly many pairs of cause and effect and countlessly many contexts to weigh up. So, even though we have some a priori reasons to suspect that CS_{ch} is more stable than the other types of causal strength, we can actually never be sure, if this is really the case in our actual world.

To summarize: It is true that CS_{ch} has an a priori advantage over the other kinds of causal strength in being invariant across different contexts. But the fact that a higher invariance of CS_{ch} is possible, or even probable, does not mean that it is actual. Whether CS_{ch} really is more invariant across different contexts and therefore more useful for causal predictions than the other kinds of causal strength is ultimately an empirical question that can probably never be settled conclusively. This is why I think that the problem at hand should be tackled from a different angle. Instead of asking which kind of causal strength is actually more invariant across different contexts, we should ask which kind of causal strength humans assume to be stable when making predictions. When we ask the second question, we are not trying to find

[21] An exception is of course the extreme case, in which $P(E|\neg C) = 1$. In that case the measure is undefined. I will discuss this issue later.

out which concept of causal strength really is more useful for predictions, instead we ask which concept of causal strength human beings assume to be more useful for predictions. It is therefore not a question about the normative superiority of a concept of causal strength, but about its descriptive superiority.

4.4.3 An Argument for Descriptive Superiority

Cheng and her colleagues have done several experiments to illustrate the descriptive superiority of CS_{ch} over alternative concepts of causal strength when it comes to capturing what human beings consider to be the intrinsic causal power of a cause on its effect. Let us first have a closer look at an experiment by Liljeholm and Cheng (2007), that aims "to identify the property of a causal relation that is assumed to be invariant, and hence generalized across contexts" (Liljeholm and Cheng, 2007, p. 1017). The authors again only compare CS_{ch} and CS_e, but, as we will see, the experiments also allow conclusions about the descriptive accuracy of CS_s, CS_l, and CS_g.[22]

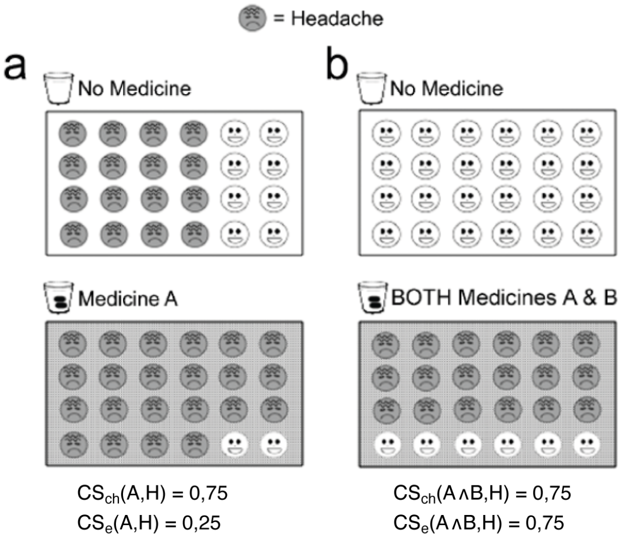


Figure 4.3: The two trials a and b for group 1. From (Liljeholm and Cheng, 2007, p. 1015).

Liljeholm and Cheng (2007) separated participants into two groups. Participants of both groups were presented with the following story (Liljeholm and Cheng, 2007, cf. p. 1018):

> "A pharmaceutical company is investigating if two allergy medicines (Medicines A and B) might produce headache as a side effect. The company has conducted two experiments that test the influence of these medicines, and you will see the results from both experiments. The two experiments were conducted in different labs, so the number of allergy patients who have a headache before receiving any medicine may vary across experiments. After reviewing the results from both experiments, you will be asked about the influence of the medicines on headache".

[22] In the following, we do not need to consider the values of CS_g separately. All the following experiments deal with examples in which causal strength is determined in randomized controlled trials, in which $P(A) = P(\neg A) = 0.5$, so $CS_g(A, H) = 4 \times P(A) \times P(\neg A) \times (P(H|A) - P(H|\neg A)) = P(H|A) - P(H|\neg A) = CS_e(A, H)$. CS_g will therefore perform just as well as CS_e in all the following experiments.

Figure 4.3 illustrates the results of the two medical trials a and b that were presented to group 1. Trial a enables to assess the strength of the causal influence that A has on headaches (H) and trial b enables to assess the strength of the causal influence that the combination of A and B has on H. $CS_{ch}(A, H)$ has the same value in trial a as $CS_{ch}(A \wedge B, H)$ has in trial b, namely 0.75. But the value of $CS_e(A, H) = 0.25$ ($CS_s(A, H) = 0.125$, $CS_l(A, H) = {}^3/_{11}$) in trial a differs from the value of $CS_e(A \wedge B, H) = 0.75$ ($CS_s(A \wedge B, H) = 0.375$, $CS_l(A \wedge B, H) = 1$) in trial b. Participants of Group 2 were presented with different results, which are summarized in Figure 4.4. This time the value of $CS_e(A, H) = 0.25$ ($CS_s(A, H) = 0.125$) in trial a is the same as the value of $CS_e(A \wedge B, H) = 0.25$ ($CS_s(A \wedge B, H) = 0.125$) in trial b. But the value of $CS_{ch}(A, H) = 0.25$ ($CS_l(A, H) = 1$) in trial a differs from the value of $CS_{ch}(A \wedge B, H) = 0.75$ ($CS_l(A \wedge B, H) = {}^3/_{11}$) in trial b.

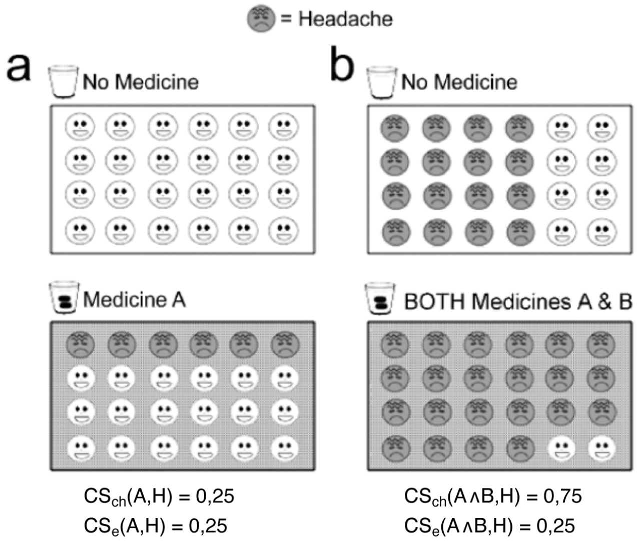


Figure 4.4: The two trials a and b for group 2. From (Liljeholm and Cheng, 2007, p. 1017).

Now, the participants of both groups were asked (Liljeholm and Cheng, 2007, cf. p. 1018):

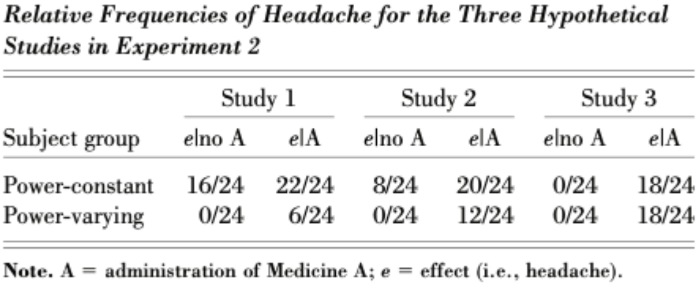
> "Based on the information from BOTH experiments, what is your best bet on whether or not Medicine B causes headache?"

Let us first try to approach this question from the point of view of a member of group 1. If one assumes that $CS_{ch}(A, H)$ measures the intrinsic causal strength of A on H, then, due to the invariance of intrinsic causal power, $CS_{ch}(A, H)$ has the same value in trial b as it has in trial a, namely 0.75. Since, for group 1, $CS_{ch}(A \wedge B, H)$ is also 0.75 in trial b, the experiments appear as evidence that B does not have any causal influence on H. The intrinsic causal power of $A \wedge B$ on H is just the same as the intrinsic causal power of A on H. So B seems to have no causal influence on H at all. Now imagine, one assumes that $CS_e(A, H)$ measures the intrinsic causal power of A on H. Then, due to the invariance of intrinsic causal power, $CS_e(A, H)$ has the same value in trial b as it has in trial a, namely 0.25. Trial b then appears as evidence that B does have a significant causal influence on H since the intrinsic causal power of the combination $A \wedge B$ on H is significantly bigger (0.75) than the intrinsic causal power of A on H. The same holds for CS_s and CS_l.

For group 2 we have the same results in reverse. If one assumes that $CS_e(A, H)$ (or CS_s) measures the intrinsic causal power of A on H, then the trials appear as evidence that B does not have any causal influence on H. But if one assumes that CS_{ch} (or CS_l) measures the intrinsic causal power of A on H, then the two trials appear as evidence that B does have a causal influence on H. So all in all, the assessment of whether the two trials appear as evidence that B has a causal influence on H depends on which kind of causal strength is considered to be intrinsic.

Now, here are the results of the survey: In group 1, 18 out of 25 responded that Medicine B has no causal influence on H. In group 2, 5 out of 25 gave the same answer. Although the number of participants is not that big, the results from both groups clearly indicate that the majority of participants by default assumed $CS_{ch}(A, H)$ to be intrinsic and therefore invariant across different contexts and not $CS_e(A, H)$ or $CS_s(A, H)$. The results present mixed evidence for the thesis that $CS_l(A, H)$ is assumed to be intrinsic. While the results from group 2 can be explained by the thesis that $CS_l(A, H)$ is assumed to be intrinsic by the majority, the results from group 1 clearly speak against this thesis.

Another experiment that indicates that humans by default consider CS_{ch} to be intrinsic was made by Buehner et al. (2003).[23] The authors present participants with several trials, in which the causal influence of different medicines on headaches is tested. In most trials the value of CS_{ch} differed from the value of CS_e. The participants were asked to determine how often a tested medicine would produce a headache, if it would be given to 100 patients that do not have a headache so far. The results clearly indicate that most participants of the experiment extrapolate the value of CS_{ch} to the hypothetical trial and therefore assume CS_{ch} to remain invariant. In yet another experiment, Liljeholm and Cheng (2007) make use of the fact that the interpretation of the concept of intrinsic causal power yields a corresponding interpretation of the concept of causal interaction. They again separate the participants of their experiment into two groups, a "power-constant" and a "power-varying group", and present each group with the results of three clinical trials, in which the causal influence of medicine A on headache H is tested. Figure 4.5 summarizes which results are shown to which group:



Relative Frequencies of Headache for the Three Hypothetical Studies in Experiment 2

	Study 1		Study 2		Study 3	
Subject group	*e*|no A	*e*|A	*e*|no A	*e*|A	*e*|no A	*e*|A
Power-constant	16/24	22/24	8/24	20/24	0/24	18/24
Power-varying	0/24	6/24	0/24	12/24	0/24	18/24

Note. A = administration of Medicine A; *e* = effect (i.e., headache).

Figure 4.5: The three studies presented to the two groups. From (Liljeholm and Cheng, 2007, p. 1019).

Again, Liljeholm and Cheng (2007) only compare the values of CS_{ch} and CS_e. But we can simply amend the results for the other measures and compare their predictions with the empirical results. For the power-constant group, $CS_{ch}(A, H)$ remains stable in all three trials, while

[23] Experiment 2 in (Buehner et al., 2003).

$CS_e(A,H)$ varies in each trial, and so do $CS_s(A,H)$ and $CS_l(A,H)$. For the power-varying group, $CS_{ch}(A,H)$, $CS_e(A,H)$ and $CS_s(A,H)$ vary across all three trials, but $CS_l(A,H)$ remains constant. Now, the participants were asked the following question (Liljeholm and Cheng, 2007, p. 1019):

> "Based on the results from ALL THREE experiments, do you think that Medicine A interacts with some factor that varies across experiments, or do you think that the medicine influences the patients in different experiments in the same way? YES: I think that the medicine interacts with some factor that varies across experiments. NO: I think that the medicine has the same influence across experiments."

For someone who assumes $CS_{ch}(A,H)$ to be intrinsic and therefore invariant by default, the results of the three trials presented to the power-constant group appear as evidence that no other factors interact with the causal influence of A on E. The results of the three trials presented to the power-varying group on the other hand would appear as evidence that other factors do interact with A's causal influence on E. But for someone, who assumes $CS_e(A,H)$, $CS_s(A,H)$ or $CS_l(A,H)$ to be intrinsic and therefore invariant by default, the results of the three trials presented to the power-constant group appear as evidence that other factors do interact with A's causal influence on E. The results of the three trials as presented to the power-varying group also appear as evidence that other factors interact with A's causal influence on E for someone, who assumes $CS_e(A,H)$ or $CS_s(A,H)$ to be intrinsic. For someone, who assumes $CS_l(A,H)$ to be intrinsic, the results of the three trials as presented to the power-varying group appear as evidence that other factors do not interact with A's causal influence on E.

Here are the results: Only 5 out of 15 in the power-constant group responded "yes", but 13 out of 15 in the power-varying group responded "yes". So, the majority answers in accord with the assumption that $CS_{ch}(A,H)$ is intrinsic. Notice that, while CS_l received mixed results in the first experiment, CS_l performs especially bad in the second experiment, since also the results of the power-varying group clearly contradict the predictions based on the assumption that $CS_l(A,H)$ is intrinsic.

To summarize: the results of the three mentioned experiments clearly indicate that humans, by default, consider causal strength as measured by CS_{ch} to be invariant across different contexts, rather than causal strength as measured by the other measures. So, CS_{ch} clearly seems to describe more accurately how humans actually evaluate the intrinsic causal power of a cause than any of the alternative concepts of causal strength mentioned above. As already pointed out, this descriptive superiority of CS_{ch} as the intrinsic causal power can be seen as an indication for the normative superiority of CS_{ch} as the intrinsic causal power. Humans are pretty successful in making causal predictions. The fact that humans employ CS_{ch} as a measure of intrinsic causal power therefore indicates that CS_{ch} is indeed the best possible choice for a measure of intrinsic causal power.

4.5 The Value of the Other Measures

It is the attribution of intrinsicness to CS_{ch} that makes the measure so fruitful for predictions, decision making, and target-oriented manipulations. But this does not mean, that the other concepts of causal strength are entirely useless. Depending on our goals, they can still provide us with very useful information. $CS_e(C,E)$, for example, gives us the *absolute* difference that the introduction of C makes for the probability of E in a given situation compared to the introduction of $\neg C$. This is in contrast to a relative difference in terms of a probabilistic increase (or decrease) of the probability of E in comparison to the contrast situation. This is why $CS_e(C,E)$ is a so called *absolute outcome measure*, which, as we will see later, can be useful in situations, in which we aim to decide whether to realize C or $\neg C$ by an intervention.[24] Suppes' measure $CS_s(C,E)$ is also an absolute outcome measure. It gives us the absolute difference that the realization of C makes for the probability of E in a given situation in contrast to doing nothing. $CS_s(C,E)$ can therefore be a helpful measure when deciding whether to intervene on a certain context by realizing C or whether to do nothing at all. The Galton measure is at its core an absolute outcome measure just like $CS_e(C,E)$, but it additionally factors in a methodological factor: it rewards when the causal strength is determined on the basis of a population in which C-tokens and $\neg C$-tokens are equally prevelant. As Fitelson and Hitchcock (2011) point out, Lewis measure $CS_l(C,E)$ provides us, at least under certain suppositions, with the probability that C is necessary for the causal production of E, which can of course be very valuable information.

4.6 The Choice of Contrast

So far, we have made a rather significant simplification in our discussions about measures of causal strength. In all but one measure that we have considered in the previous sections,[25] the causal strength of an event C on another event E is measured with regard to a contrasting event, which is an alternative to C. So far, we have simply denoted this event by $\neg C$ and we have presupposed that this choice is unambiguous. But typically, this is not the case. Consider our examples from above, in which we aimed to measure the causal strength of the event A (taking medicine A) on the event H (having a headache). We have simply presupposed that $\neg A$ represents the event *taking no medicine at all*. But clearly this is not the only alternative to A. $\neg A$, the negation of *taking medicine A*, can be realized in many different ways. For example, by A' = *taking medicine B instead of A*, A'' = *taking no medicine at all*, or A''' = *hitting ones head with a hammer instead of taking medicine A*, and so on. The problem becomes even more obvious with factors that come in degrees. We might, for example, want to know the causal strength of the event C = *taking 200 mg of substance A* on the event H = *the patient suffers from headache*. Now, what is the correct choice for the alternative $\neg C$? Is it C' = *taking* 199 *mg of substance A*, C'' = *taking* 100 *mg of substance A*, C''' = *taking* 0 *mg of substance A*, or C'''' = *taking* 1 *kg of substance A*?[26]

[24] Stegenga (2015) and Sprenger and Stegenga (2017) actually make the argument, that in certain situations of decision making, we should avoid using relative outcome measures, like $CS_{ch}(C,E)$, and only use absolute outcome measures like $CS_e(C,E)$. This argument does not contest the assumption that $CS_{ch}(C,E)$ is the best candidate for measuring the intrinsic causal power of C on E. It only contests the fruitfulness of relative outcome measures for certain situations of decision making, especially in comparison with absolute outcome measures. This is why I will deal with this argument later in chapter 5.

[25] The exception is Suppes' measure.

[26] In the framework of causal models, the problem can be formulated like this: When C has more than two values and we want to measure the causal strength of an event $C = c$ on the effect $E = e$, which value c' of C with $c' \neq c$ should be the one that represents the contrast to $C = c$ that appears in the measure of causal

Clearly, the choice of contrast can make a significant difference to the results of a measure of causal strength. Imagine, for example, that we have four groups, each consisting of 1000 randomly chosen patients. In group 1, every patient takes 0 mg of substance A (C') and it turns out that 50 out of 1000 have a hedeache. In group 2, every patient takes 100 mg of substance A (C'') and it turns out that 200 out of 1000 have a hedeache. In group 3, every patient takes 200 mg of substance A (C) and it turns out that 400 out of 1000 have a hedeache. And in group 3, every patient takes 199 mg of substance A (C''') and it turns out that 400 out of 1000 have a hedeache. Now, we want to use CS_{ch} to determine the causal strength of $C =$ *taking 200 mg of substance A* on $E =$ *the patient suffers from headache.* Since we have 3 possible contrasts to C, we have the following results:

$$C'\colon\ CS_{ch}(C,E) = \frac{P(E|C) - P(E|C')}{1 - P(E|C')} = \frac{0.4 - 0.05}{0.95} = 0.37$$

$$C''\colon\ CS_{ch}(C,E) = \frac{P(E|C) - P(E|C'')}{1 - P(E|C'')} = \frac{0.4 - 0.2}{0.8} = 0.25$$

$$C'''\colon\ CS_{ch}(C,E) = \frac{P(E|C) - P(E|C''')}{1 - P(E|C''')} = \frac{0.4 - 0.4}{0.6} = 0$$

This shows, that if we want to determine the intrinsic causal power of C on E, which is a definite value that remains invariant across different contexts, then we cannot leave the choice of the contrast C', which appears in the measure $CS_{ch}(C,E)$, up to chance.

Luckily, the presuppositions made for the deduction of Cheng's measure already give us a clear guidance for choosing adequate contrasts. In assumption CP1, we have assumed that there are only two factors that can causally influence the effect E, namley the cause candidate C itself and the background causes U. Whether we realize C or the contrast C' by an intervention should not make any difference to U or the probability of U (CP3). Now, imagine that the realization of C' is an event that does have a non-zero amount of causal power on E. In that case, there would be an additional factor which has a non-zero amount of causal power on E, that is present when $P(E|C')$ is measured, but that is not present when $P(E|C)$ is measured. So, U and the probability of U would have changed in both contexts. The presuppositions that we have used for the deduction of Cheng's measure $CS_{ch}(C,E)$ as a measure of the intrinsic causal power $g_{C,E}$ of C on E are therefore not fulfilled. So, if we choose a contrast C' to C that does have a causal influence on E itself, then we have to be aware that the formula

$$\frac{P(E|C) - P(E|C')}{1 - P(E|C')} \tag{4.10}$$

does not provide us with the intrinsic causal power of C on E. This gives us a clear guideline for choosing an adequate contrast to C, if we want to measure the intrinsic causal power of C on E by using formula (4.10): Take only alternatives C' to C that do not themselves have any causal influence on E. With this choice, the presuppositions made for the deduction of formula (4.10) as a measure of intrinsic causal power of C on E are satisfied and we can therefore use it for this purpose.[27]

strength?

[27] There may be situations, in which every possible alternative to C has a causal influence on the effect in

Notice that compliance with this instruction is only required, if we aim for measuring the intrinsic causal power of C on E with formula (4.10). Otherwise, especially for the other measures that determine other types of causal strength, there are no restrictions for choosing an alternative C' to C.[28] If we, for example, want to determine $CS_e(C, E)$, then there is nothing that restricts our choice of the contrast C'. Remember that $CS_e(C, E)$ provides us with the absolute difference that the realization of C makes for the probability of E in a given situation compared to the realization of C' in that situation. Whenever we want to decide between the alternative actions of realizing C or realizing C', $CS_e(C, E) = P(E|C) - P(E|C')$ provides us with useful information, no matter if C' does have a causal influence on E itself or not. We only have to be aware of the fact, that the value of $CS_e(C, E)$ depends on the choice of C'. So, whenever there are several alternatives to C that yield different results in a measure of causal strength $CS(C, E)$, it makes sense to treat $CS(C, E)$ as a three-place relation. We should not only talk about the causal strength $CS(C, E)$ of C on E but about the causal strength $CS(C|C', E)$ of C on E relative to C'.[29] So, in the following, whenever I simply speak of the causal strength $CS(C, E)$ of C on E, I presuppose that all alternatives to C that are admissible in the measure $CS(C, E)$ yield the same value of $CS(C, E)$.

4.7 Sprenger's Arguments Against Cheng's Measure

We have seen that there are several different conceptions of causal strength. Which kind of causal strength provides us with the most useful information largely depends on the goals and purposes in the given situation. But we have also seen that human beings typically ascribe Cheng's concept of causal stregth a special role, since they identify it with the intrinsic causal power, that is supposed to remain invariant across contextual changes. It is this assumption of invariance that makes it an indispensable tool in the highly effective toolbox of human causal reasoning. But despite all the indications to the special importance of Cheng's measure CS_{ch}, Jan Sprenger (2018) has put forward some arguments that not only challenge the special status of CS_{ch}, but even call into question whether CS_{ch} is an adequate measure of causal strength at all. With a few formal assumptions about probabilistic measures of causal strength, Sprenger (2018) shows that only CS_e, and all measures that are ordinal equivalent to CS_e, satisfies a group of adequacy conditions, that he considers to be highly desirable. Motivated by these results, Sprenger adopts a monist position concerning the variety of causal strength measures: All measures that are not ordinal equivalent to CS_e are inadequate and should therefore be disposed of. So, if Sprenger is right, the adequate position for CS_{ch} would not be a bedrock of causal reasoning, but the dustbin of history.

question. In that case our just formulated guideline is of no help. But we will later learn an alternative method for determining the intrinsic causal power $g_{C,E}$, which can be applied in such a situation.

[28] We have mentioned that it is an important presupposition for all measures of causal strength that background causes remain constant in both contexts in which $P(E|C)$ and $P(E|C')$ are measured. But this presupposition does not include that C' must not have any causal influence on E itself.

[29] Notice that this is not necessary for $CS_{ch}(C, E)$, since when determining the intrinsic causal power of C on E, all alternatives to C, that are admissible for the deployment in formula (4.10), always yield the same result.

Sprenger (2018) starts with the following adequacy constraint for any measure of causal strength:[30]

Generalized Difference Making (GDM). *There is a real-valued, continuous function f : $[0,1]^2 \mapsto \mathbb{R}$ such that for any putative cause C and any putative effect E, there exists an alternative value C' to C such that for the causal strength $CS(C,E)$ of C on E, we have:*

$$CS(C,E) = f(P(E|C), P(E|C')) \tag{4.11}$$

where f is non-decreasing in the first argument and non-increasing in the second argument.

Sprenger characterizes GDM as a "very general" (Sprenger and Hartmann, 2019, p. 160) and therefore unassuming adequacy constraint, which any measure of causal strength that has been proposed in the literature so far, especially those mentioned in table 4.1, seems to satisfy.[31] As he points out, it is based on the intuition that the causal strength of C on E is the higher, the more probable E gets when evoking C, and the less probable E gets when evoking C' and that nothing else influences the assessment of causal power of C on E.

Sprenger is now able to show, that only measures of the form $CS(C,E) = P(E|C) - P(E|C')$ can satisfy both GDM and the following adequacy constraint:[32]

Seperability of Effects (SE). *Let $\tilde{E}$ be an abstract event with two mutually exclusive and collectively exhaustive determinates E and E' of $\tilde{E}$. If we use the same contrast C' for a cause candidate C in the determination of $CS(C,\tilde{E})$, $CS(C,E)$, and $CS(C,E')$, then we have:*

$$\begin{aligned}
&CS(C,\tilde{E}) > CS(C,E) \text{ \textit{iff} } CS(C,E') > 0\\
&CS(C,\tilde{E}) = CS(C,E) \text{ \textit{iff} } CS(C,E') = 0\\
&CS(C,\tilde{E}) < CS(C,E) \text{ \textit{iff} } CS(C,E') < 0
\end{aligned} \tag{4.12}$$

SE describes how the causal strength of a cause candidate C on an event $\tilde{E}$ is connected to C's causal strength on two more specific realizations of $\tilde{E}$.[33] Consider the following example. A student has to write an exam that she can either pass or fail. But there are two different ways of passing the exam. A passed exam is graded either with a plus or with a minus. Passing the

[30] The following formulation of GDM is actually from (Sprenger and Hartmann, 2019, p. 160), where an amended version of (Sprenger, 2018) is republished. In (Sprenger, 2018), GDM contains a different order of the quantifiers such that GDM allowed for the function f to change when the causal setting, the putative cause or the putative effect under consideration change. But as becomes clear in the proofs of the theorems in (Sprenger, 2018), this was unintentional. The reformulation from (Sprenger and Hartmann, 2019, p. 160), which is presented here, is clearly the intended formulation of GDM. In the following, any direct quote from Sprenger's (2018) article will be taken from the republication in (Sprenger and Hartmann, 2019). Notice also that Sprenger uses a slightly different notation than I do here. He uses the framework of probabilistic causal models and considers $P(E|C)$, for example, to be the probability $\mathcal{P}_{do(C)}(E)$ of E after realizing C by an intervention, which precludes the possibility of confounding. As pointed out above, I also presuppose that for the determination of $P(E|C)$ and $P(E|C')$ any confounding factors are controlled for. The notational differences are therefore irrelevant for our discussion.

[31] Notice that C's alternative C' in GDM does not have to be $\neg C$. Sprenger also admits $P(E)$, where we conditionalize on nothing at all, as an instance of $P(E|C')$, which is why Suppes' measure also satisfies GDM.

[32] See (Sprenger and Hartmann, 2019, p. 163 f.). My formulation of SE slightly differs from Sprenger's formulation, but without distorting any of its substantial meaning.

[33] In chapter 9, I will deal in more detail with causation on different supervenience levels.

exam is the abstract event $\tilde{E}$ that has two more specific determinates: $E =$ *passing with a plus* and $E' =$ *passing with a minus*. Take $C =$ *studying the the day before the exam* as the cause candidate with $C' =$ *not studying the day before the exam* as its contrast. Now, imagine that we already know that C has a certain amount of causal strength on $\tilde{E} =$ *passing the exam*. SE now tells us that the causal strength of C on $\tilde{E} =$ *passing the exam* is bigger than (smaller than / equal to) the causal strength of C on $E =$ *passing with a plus* if and only if the causal strength of C on $E' =$ *passing with a minus* is bigger than (smaller than / equal to) 0.[34]

I agree with Sprenger that SE is indeed a very intuitive adequacy constraint for a measure of causal strength. The result of Sprenger's theorem therefore seems to be good news for CS_e and CS_s, which both conform to the form $CS(C, E) = P(E|C) - P(E|C')$, and bad news for CS_{ch}, which does not correspond to this form and therefore cannot satisfy both GDM and SE. But Sprenger has even more to offer. Sprenger proves that only a causal strength measure $CS(C, E)$ that is ordinal equivalent[35] to $CS_e(C, E) = P(E|C) - P(E|\neg C)$ can satisfy both GDM with the choice of $\neg C$ for C' and the following adequacy condition of Multiplicativity:[36]

Multiplicativity. *If C has causal strength $CS(C, X)$ on X, X has causal strength $CS(X, E)$*[37] *on E, and C causally influences E only via X, then:*

$$CS(C, E) = CS(C, X) \times CS(X, E) \tag{4.13}$$

For a causal path from C to E with several mediators X_1, ..., X_n, where each link satisfies the conditions of Multiplicativity, Multiplicativity implies:

$$CS(C, E) = CS(C, X_1) \times ... \times CS(X_n, E) \tag{4.14}$$

As Sprenger (2018) points out, Multiplicativity has several intuitive consequences when it comes to indirect causation, for example, that the causal strength of C on E, which goes via the mediators X_1, ..., X_n, cannot be higher than the causal strength between any succeeding mediators from C to E. It also follows from Multiplicativity that any causal link between C and E, which has maximal causal strength, does not weaken the ultimate causal strength of C on E. Sprenger's result therefore appears to be another corroboration of Eells' measure of causal strength and a clear rebuttal of CS_{Ch}, which is not ordinal equivalent to CS_e. But in the following sections, I will show that this appearance is deceptive. Instead of rebutting Cheng's concept of causal power, Sprenger's arguments only highlight that the scope of Cheng's Power PC theory is still too restrictive and needs to be expanded. I will therefore use Sprenger's arguments as a starting point to tackle this weakness and to generalize Cheng's Power PC theory and thereby the applicability of her measure of causal strength.[38]

[34] A similar example is given in (Sprenger and Hartmann, 2019, p. 163).

[35] The causal measures CS_1 and CS_2 are ordinal equivalent if and only if for all pairs (C_1, E_1) and (C_2, E_2) it holds that: $CS_1(C_1, E_1) \lesseqqgtr CS_1(C_2, E_2)$ if and only if $CS_2(C_1, E_1) \lesseqqgtr CS_2(C_2, E_2)$.

[36] See (Sprenger and Hartmann, 2019, p. 167).

[37] Note that I am assuming here that all alternatives to X that are admissible in the measure $CS(X, E)$ yield the same value of $CS(X, E)$.

[38] I do not deal with Sprenger's (2018) third argument for CS_e, in which he aims to show that CS_{Ch} satisfies a property which he calls 'No Dilution for Irrelevant Effects (Prevention)'. My rebuttal of the two arguments presented here, will also provide a rebuttal of this third argument, since it is based on the same problematic

4.8 The Power PC Theory Expanded

Sprenger's arguments seem to put CS_{ch} in a precarious position. I have already pointed out that I agree with Sprenger that Seperability of Effects and Multiplicativity are indeed very intuitive adequacy constraints for any concept of causal strength. Also, Sprenger's results leave no doubt that the formula

$$f_{ch}(C, E) = \frac{P(E|C) - P(E|\neg C)}{1 - P(E|\neg C)} \tag{4.15}$$

which clearly satisfies GDM, cannot also satisfy Seperability of Effects and Multiplicativity. But, and this is the essential point, we should not generally identify the formula f_{ch} with Cheng's concept of causal power CS_{ch}.[39] Cheng (1997) only showed that the value of $f_{ch}(C, E)$ is equal to the value of $CS_{ch}(C, E) = P(C \text{ produces } E|C)$, if C and E appear in a Cheng-scenario as depicted in figure 4.6.

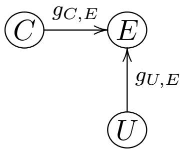


Figure 4.6: A Cheng-scenario.

But we do not yet know, how values of $CS_{ch}(C, E) = P(C \text{ produces } E|C)$ can be determined in more complex causal scenarios that do not satisfy conditions CP1-CP3. This is why the fact that f_{ch} violates Multiplicativity and SE does not imply that CS_{ch} violates Multiplicativity and SE. For assessing, whether CS_{ch} violates or satisfies Multiplicativity and SE, we first have to figure out, how values of CS_{ch} can be determined in more complex causal scenarios, where Multiplicativity and Seperability of Effects do play a role.

4.8.1 Multiplicativity: A First Amendment of the Power PC Theory

The condition of Multiplicativity concerns causal scenarios with indirect causation. Consider the example shown in figure 4.7, in which we have C with some generative causal power on X, X with some generative causal power on E, and U with some generative causal power on E.

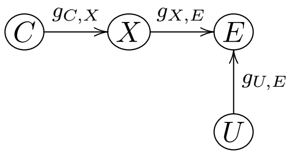


Figure 4.7: Indirect causation.

assumption. But more crucially, I do not think that this third argument even really needs a rebuttal, since I do not consider the property 'Dilution for Irrelevant Effects (Prevention)' to be problematic in the first place.

[39] This is why I use 'f_{ch}' from now on to refer to the formula in equation (4.15). 'CS_{ch}', on the other hand, denotes Cheng's concept of causal power as it is characterized in axiom 1 of the power PC theory, namely as $P(C \text{ produces } E|C)$, and that aims to explicate the concept of intrinsic causal power.

Cheng (1997) deduced that $f_{ch}(C, E) = CS_{ch}(C, E)$ if we are dealing with a causal scenario that satisfies conditions CP1-CP3. But the given causal scenario of indirect causation violates CP1 and CP3, since there are three distinct generative causes of E, two of which are not probabilistically independent. It is therefore not yet clear, how the value of $CS_{ch}(C, E)$ can be determined in the given scenario. But our ignorance goes even deeper. Axiom C1 characterizes the generative causal power $g_{C,E}$ of C on E as the probability that C produces E given that C is the case. The concept of causal production employed in this characterization is a theoretical term that gets its meaning only from the axioms C2-C5. In the given scenario of indirect causation, we assume that C has some generative causal power on X and X has some generative causal power on E. This means that C has the potential to causally produce X and X has the potential to causally produce E. But this leaves completely open, whether C also has the potential to causally produce E. The existing axioms of Cheng's Power PC theory do not enable us to deduce anything about the causal relationship between C and E in the given scenario of indirect causation. To clarify the causal relationship between C and E, given that C can causally produce X and X can causally produce E, we need to add another axiom to the Power PC theory, that is able to capture our intuitions about cases of indirect causation. Here is my proposal:

C6 (Transitivity) If C produces X and X produces E, then C produces E.

I do not only consider C6 to be highly intuitive, the condition also conforms with several philosophical explications of the concept of causal production.[40] With this additional axiom at hand, we can show the following:

Multiplicativity of CS_{ch}. *Let C be a generative cause of X with causal power $g_{C,X}$. Let X be a generative cause of E with causal power $g_{X,E}$. Let C's only causal influence on E go via X and let X satisfy the following condition:*

CP4 Whether X produces E is independent of how (by which cause) X itself is produced, given that X is produced.

Then the generative causal power $g_{C,E}$ of C on E is given by:

$$g_{C,E} = g_{C,X} \times g_{X,E} \tag{4.16}$$

Proof. We have $g_{C,E} = P(C \text{ produces } E|C)$ by C1. We then have $P(C \text{ produces } E|C) = P(C \text{ produces } X \wedge X \text{ produces } E|C)$ by axiom C6. Next, we have $P(C \text{ produces } X \wedge X \text{ produces } E|C) = P(C \text{ produces } X|C) \times P(X \text{ produces } E|C \text{ produces } X \wedge C)$ by probability theory. With axioms C2 and C3, X follows from C produces $X \wedge C$, so we have $P(C \text{ produces } X|C) \times P(X \text{ produces } E|C \text{ produces } X \wedge C) = P(C \text{ produces } X|C) \times P(X \text{ produces } E|X \wedge C \text{ produces } X \wedge C)$. With CP4, we get: $P(C \text{ produces } X|C) \times P(X \text{ produces } E|X \wedge C \text{ produces } X \wedge C) = P(C \text{ produces } X|C) \times P(X \text{ produces } E|X) = g_{C,X} \times g_{X,E}$. □

[40] See, for example, (Hall, 2004) and (Dowe, 2000). I will discuss philosophical accounts of causal production in Chapter 6. Notice that there are clear counterexamples against the transitivity of actual causation. But, as we will also clarify in more detail later, actual causation is not the same as causal production. There are actual causes that are no causal producers and it is only those actual causes that violate transitivity.

This shows, if we amend Cheng's Power PC theory with axiom C6, we can show that the resulting concept of causal power, which I still denote by CS_{ch}, does indeed satisfy Multiplicativity.[41] But how is this compatible with Sprenger's result that any measure of causal strength that satisfies GDM and Multiplicativity must be ordinal equivalent to CS_e? How can CS_{ch} satisfy Multiplicativity while not being ordinal equivalent to CS_e? The answer is of course that CS_{ch} does not satisfy GDM. Sprenger assumes that CS_{ch} satisfies GDM, because he identifies CS_{ch} with f_{ch}. But, as we will now see, it is wrong to assume that f_{ch} can always be used to determine the value of CS_{ch}.

4.8.2 Where CS_{ch} and f_{ch} Come Apart

There are many causal scenarios, in which the formula $f_{ch}(C, E)$ can be used to determine the causal power $CS_{ch}(C, E)$ of a cause C on its effect E. This includes the very basic causal scenarios that constitute the original scope of application of Cheng's Power PC theory, namely the scenarios that satisfy conditions CP1-CP3. But there are also many other causal scenarios, in which the formula $f_{ch}(C, E)$ does not provide us with the value of $CS_{ch}(C, E)$. To illustrate, how the values of $f_{ch}(C, E)$ and $CS_{ch}(C, E)$ come apart, let us consider two examples.

The first example as shown in figure 4.8 is a simple scenario of indirect causation, in which C is the only generative cause of X, in which any pair of generative causes satisfies the independence assumptions of CP2 and CP3, and in which the mediator X satisfies CP4.

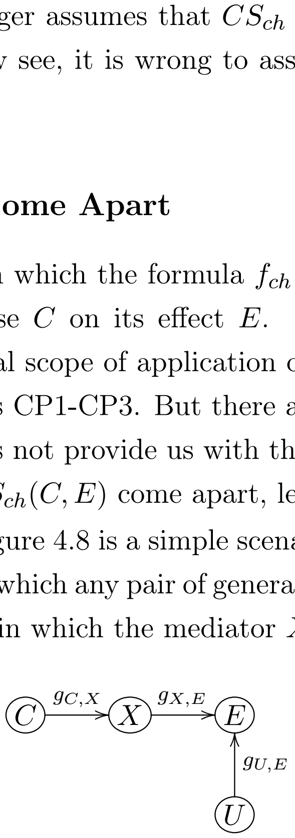


Figure 4.8: Indirect causation.

We have already shown that the causal power $g_{C,E}$ of C on E is given by $g_{C,X} \times g_{X,E}$. But we can also determine $g_{C,E}$ by employing the formula $f_{ch}(C, E)$. To see this, notice that the probability of E is given by the following formula:

$$P(E) = P(X) \times g_{X,E} + P(U) \times g_{U,E} - P(X) \times g_{X,E} \times P(U) \times g_{U,E} \tag{4.17}$$

Since we assume that C is the only generative cause of X in the given scenario, the probability of X is given by: $P(X) = P(C) \times g_{C,E}$. This gives us:

$$P(E) = P(C) \times g_{C,E} \times g_{X,E} + P(U) \times g_{U,E} - P(C) \times g_{C,E} \times g_{X,E} \times P(U) \times g_{U,E} \tag{4.18}$$

Conditionalizing on C gives us:

[41] At least as long as the mediator satisfies condition CP4, which I consider to be a reasonable restriction. But notice that, even if CP4 is not fulfilled and the causal power $g_{X,E}$ of X on E is sensitive to the producer of X, a certain form of Multiplicativity still holds for CS_{ch}, namely: $g_{C,E} = g_{C,X} \times g^{C}_{X,E}$, where $g^{C}_{X,E}$ is the generative causal power that X has on E, whenever X is produced by C.

$$P(E|C) = g_{C,E} \times g_{X,E} + P(U) \times g_{U,E} - g_{C,E} \times g_{X,E} \times P(U) \times g_{U,E} \tag{4.19}$$

Conditionalizing on $\neg C$ gives us:

$$P(E|\neg C) = P(U) \times g_{U,E} \tag{4.20}$$

We therefore get:

$$\begin{aligned} &P(E|C) - P(E|\neg C) = \\ &= g_{C,E} \times g_{X,E} - g_{C,E} \times g_{X,E} \times P(U) \times g_{U,E} \\ &= g_{C,E} \times g_{X,E} \times (1 - P(U) \times g_{U,E}) \end{aligned} \tag{4.21}$$

Since $1 - P(U) \times g_{U,E} = 1 - P(E|\neg C)$, we get:

$$f_{ch} = \frac{P(E|C) - P(E|\neg C)}{1 - P(E|\neg C)} = g_{C,E} \times g_{X,E} \tag{4.22}$$

This shows, in the given example of indirect causation, we can use the formula $f_{ch}(C, E)$ to determine the value of $CS_{ch}(C, E)$. But let us now consider a causal scenario as shown in figure 4.9, which is just like the previous one, with the only difference that U is not a direct cause of E, but it influences E also via the mediator X:

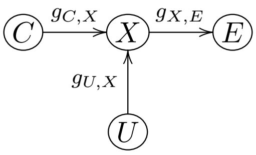


Figure 4.9: Indirect causation with an externally influenced mediator.

Here again, we have already shown that the causal power $g_{C,E}$ of C on E is given by $g_{C,X} \times g_{X,E}$. But this time, we cannot use $f_{ch}(C, E)$ to determine $g_{C,E}$. To see this, consider that the probability of E is given by the following formula:

$$P(E) = P(X) \times g_{X,E} \tag{4.23}$$

The probability of X is given by:

$$P(X) = P(C) \times g_{C,X} + P(U) \times g_{U,X} - P(C) \times g_{C,X} \times P(U) \times g_{U,X} \tag{4.24}$$

So, we get:

$$P(E) = (P(C) \times g_{C,X} + P(U) \times g_{U,X} - P(C) \times g_{C,X} \times P(U) \times g_{U,X}) \times g_{X,E} \tag{4.25}$$

Conditioning on C gives us:

$$P(E|C) = (g_{C,X} + P(U) \times g_{U,X} - g_{C,X} \times P(U) \times g_{U,X}) \times g_{X,E} \tag{4.26}$$

Conditioning on $\neg C$ gives us:

$$P(E|\neg C) = P(U) \times g_{U,X} \times g_{X,E} \tag{4.27}$$

Therefore:

$$\begin{aligned}
&P(E|C) - P(E|\neg C) = \\
&= (g_{C,X} + P(U) \times g_{U,X} - g_{C,X} \times P(U) \times g_{U,X}) \times g_{X,E} - P(U) \times g_{U,X} \times g_{X,E} \\
&= [g_{C,X} + P(U) \times g_{U,X} - g_{C,X} \times P(U) \times g_{U,X} - P(U) \times g_{U,X}] \times g_{X,E} \\
&= [g_{C,X} - g_{C,X} \times P(U) \times g_{U,X}] \times g_{X,E} \\
&= g_{C,X} \times g_{X,E} - g_{C,X} \times g_{X,E} \times P(U) \times g_{U,X} \\
&= g_{C,X} \times g_{X,E} \times (1 - P(U) \times g_{U,X})
\end{aligned} \tag{4.28}$$

This gives us:

$$g_{C,E} = g_{C,X} \times g_{X,E} = \frac{P(E|C) - P(E|\neg C)}{1 - P(U) \times g_{U,X}} = \frac{P(E|C) - P(E|\neg C)}{1 - P(X|\neg C)} \tag{4.29}$$

This shows that in our second example of indirect causation, the formula $f_{ch}(C, E)$ cannot be used to determine the value of $CS_{ch}(C, E)$. But instead of $f_{ch}(C, E)$, we can use another formula that, just like $f_{ch}(C, E)$ contains the covariation of C and E, $\Delta P_{C,E} = P(E|C) - P(E|\neg C)$, in its numerator. In the following, I will call any formula $f(C, E)$ that is used to determine the value of causal strength of C on E and that contains the term $\Delta P_{C,E} = P(E|C) - P(E|\neg C)$, a '$\Delta$-formula'. Notice though, that the Δ-formula in (4.29), which is able to determine the value of $CS_{ch}(C, E)$ in the given example, contains a term, that does not appear in $f_{ch}(C, E)$, namely $P(X|\neg C)$. The given Δ-formula does therefore not conform with GDM, which demands that the formula that determines the causal strength of C on E must be a function of only the two terms $P(E|C)$ and $P(E|C')$.

We can give the Δ-formula in (4.29) some intuitive underpinning. As argued in section 4.4, it is a central feature of Cheng's concept of causal power, CS_{ch}, that it is not susceptible to floor-effects. In general, $CS_{ch}(C, E)$ averts floor-effects by accounting for the relative frequency of cases of overdetermination, that are cases in which E is already produced by some other cause besides C, but in which C would also have produced E all on its own. Just consider a randomized controlled trial with a test-group, in which we determine the value of $P(E|C)$, and a control-group, in which we determine the value of $P(E|\neg C)$. Now imagine that there is another potential cause U of E present in both groups and U causally influences E independently from C. Since U produces a certain amount of E-tokens in both groups, it 'raises the floor'. By measuring the difference $P(E|C) - P(E|\neg C)$ we only count the number of E-tokens that are produced by C and that are not also produced by U. All the cases, in which C produces an E-token that is already produced by U are not taken into account. They are simply not measurable by the difference $P(E|C)-P(E|\neg C)$. This is why the measured difference $P(E|C)-P(E|\neg C)$ has to be put in relation to the relative frequency of the cases, in which C's production of E is measurable in the first place. This feature of CS_{ch} can easily be seen in action in a Cheng-scenario, in which $f_{ch}(C, E)$ adequately determines the value of $CS_{ch}(C, E)$:

$$f_{ch}(C, E) = \frac{P(E|C) - P(E|\neg C)}{1 - P(E|\neg C)} \tag{4.30}$$

First, $P(E|C) - P(E|\neg C)$ gives us the measurable difference that C makes for the probability of E in the actual situation. Then the denominator is there to avert the floor-effect. $P(E|\neg C)$ is the relative frequency of all cases, in which E is already produced by a cause different from C. So in all those cases, any causal influence of C on E is not measurable. This means that $1 - P(E|\neg C)$ is the relative frequency of all the cases, in which C's causal influence on E is measurable. So, $f_{ch}(C, E)$ puts the relative frequency of those cases, in which C measurably produced E-tokens, in relation to the relative frequency of those cases, where C's production of E-tokens is measurable in the first place.

But the formula $f_{ch}(C, E)$ will not do the job of averting floor effects correctly as soon as we have a case of indirect causation that includes an externally influenced mediator X between C and E. To illustrate, let us consider a simple example with the scenario from figure 4.9. Let us assume the following probabilities and values of causal power:

- $g_{C,X} = g_{U,X} = 0.8$
- $g_{X,E} = 0.5$
- $P(U) = 1$

This gives us: $P(X|C) = 0.96$, $P(E|\neg C) = 0.4$, $P(E|C) = 0.48$. We then have:

$$f_{ch}(C, E) = \frac{P(E|C) - P(E|\neg C)}{1 - P(E|\neg C)} = \frac{0.08}{0.6} = 0.1333 \tag{4.31}$$

The formula $f_{ch}(C, E)$ puts the measurable difference that C makes for E in the given situation, which is $P(E|C) - P(E|\neg C) = 0.08$, in relation to the relative frequency of cases, namely $1 - P(E|\neg C) = 0.6$, where E is not produced by U. This would avert the floor-effect correctly, if $1 - P(E|\neg C) = 0.6$ would be the relative frequency of the cases in which C's production of E is measurable. But this is not the case. In the given causal scenario, C only produces E if and only if C produces X and X produces E. But C's production of X is only measurable in 20% of all cases, since X is already caused by another cause in 80% of all cases. This means that C's production of E is also only measurable in 20% of all cases. So, while $f_{ch}(C, E)$ puts the measured difference that C makes for E in relation to $1 - P(E|\neg C) = 0.6$, it should instead have put it in relation to $1 - P(X|\neg C) = 0.2$, which is exactly what our newly deduced Δ-formula does:

$$\frac{P(E|C) - P(E|\neg C)}{1 - P(X|\neg C)} = \frac{0.08}{0.2} = 0.4 \tag{4.32}$$

Since $1 - P(E|\neg C) = 1 - P(X|\neg C) \times g_{X,E}$, using $f_{Ch}(C, E)$ to determine the value of $CS_{ch}(C, E)$ leads to an underestimation of the floor effect in the given situation. The lower the causal power

$g_{X,E}$ of X on E for a given cause U of X with $P(U) \times g_{U,X} > 0$, the more the actual height of the floor-effect will be underestimated. This results in an underestimation of the cases, in which C's production of E is not measurable due to overdetermination, which in turn results in an underestimation of C's causal power on E by the formula $f_{Ch}(C, E)$.

Let us take stock. We have shown that $CS_{ch}(C, E)$, the measure of causal power that results from our expanded Power PC theory, is not always equatable with $f_{ch}(C, E)$. It crucially depends on the causal scenario, in which the cause C and the effect E are embedded, whether $f_{ch}(C, E)$ provides us with the value of $CS_{ch}(C, E)$. We have seen an example, in which $f_{ch}(C, E)$ fails to do so and where, instead, a different Δ-formula is needed to correctly determine the value of $CS_{ch}(C, E)$. Having gained these insights, it is now time to reconsider, what Sprenger calls a "very general adequacy constraint" (Sprenger and Hartmann, 2019, p. 160), namely GDM. GDM contains two crucial claims about a given concept of causal strength. The first is that there is a single, concrete formula f that can determine the value of causal strength of a cause C on its effect E in any causal scenario. The second claim is that this formula is a function of only the two terms $P(E|C)$ and $P(E|\neg C)$. We have now seen that both these claims are wrong for $CS_{ch}(C, E)$. How exactly the Δ-formula, that correctly determines the value of $CS_{ch}(C, E)$, looks like, depends on the causal scenario, in which C and E are embedded. And at least some causal scenarios require a Δ-formula for the determination of $CS_{ch}(C, E)$, that contains other terms besides $P(E|C)$ and $P(E|\neg C)$. This shows: Cheng's concept of causal power, CS_{ch}, or, more correctly, our generalized version of it, does not satisfy GDM. This is why CS_{ch} can satisfy the condition of Multiplicativity without being ordinally equivalent to CS_e.

4.8.3 SE: A Second Amendment of the Power PC Theory

Seperability of Effects concerns causal scenarios that include a cause candidate C, an abstract effect $\tilde{E}$ and two more specific, mutually exclusive, but collectively exhaustive determinates of $\tilde{E}$, E and E'. Although, intuitively, the causal power of C on E and the causal power of C on E' are closely connected to the causal power of C on $\tilde{E}$, Cheng's Power PC theory does not account for this intuition. The existing axioms do not imply anything about how the causal power of C on E and the causal power of C on E' is related to C's causal power on $\tilde{E}$. To change this, we again need to amend Cheng's Power PC theory by incorporating another axiom. Here is my proposal:

C7 Let $\tilde{E}$ be an abtract event and let E and E' be two more specific, mutually exclusive, but collectively exhaustive determinates of $\tilde{E}$. Then: C produces $\tilde{E}$ if and only if C produces E or C produces E'.

Here again, I consider this axiom to be highly intuitive. Whenever an event C causally produces a specific event, it also causally produces every more abstract determinable of this event. And whenever an event C causally produces an event, which has several distinct determinates, then C produces some specific determinate of the event. Since E and E' are the two mutually exclusive and collectively exhaustive determinates of the event $\tilde{E}$, the two events C *produces* E and C *produces* E' are also mutually exclusive. We therefore have:

$$\begin{aligned}&P(C \text{ produces } \tilde{E}|C) = \\ &= P(C \text{ produces } E \vee E'|C) \\ &= P(C \text{ produces } E|C) + P(C \text{ produces } E'|C)\end{aligned} \tag{4.33}$$

This directly gives us the following theorem:

Summation of Effects. *If E and E' are two more specific, mutually exclusive, and collectively exhaustive determinates of an event $\tilde{E}$, and if C has a generative causal power $g_{C,E}$ on E and a generative causal power $g_{C,E'}$ on E', then C's generative causal power $g_{C,\tilde{E}}$ on $\tilde{E}$ is given by: $g_{C,\tilde{E}} = g_{C,E} + g_{C,E'}$.*

Summation of Effects clearly implies Seperability of Effects, which means that the concept of causal power, CS_{ch}, that results from our amended Power PC theory, also satisfies Seperability of Effects. Here again, CS_{ch} can satisfy SE without conforming to the form $P(E|C) - P(E|C')$, because CS_{ch} does not satisfy GDM.

4.8.4 Is GDM a Reasonable Adequacy Constraint?

Sprenger (2018) claims that only CS_e, or ordinally equivalent measures, are adequate measures of causal strength. He bases this claim on the argument that only CS_e, or ordinally equivalent measures, satisfy the two intuitive adequacy constraints of SE and Multiplicativity. For proving this, he assumes that every measure of causal strength, including CS_{ch}, conforms to GDM. I have argued in the previous sections, that this assumption is wrong. I have shown that CS_{ch}, the concept of causal strength that results from our expanded Power PC theory, does indeed satisfy Multiplicativity and SE. But it violates GDM. The intuitive adequacy of SE and Multiplicativity can therefore not be used as arguments against CS_{ch}. On the contrary, the intuitive adequacy of SE and Multiplicativity are now arguments in favor of CS_{ch}. Nonetheless, there is still a way to use Sprenger's results as arguments against CS_{ch}, namely by arguing that GDM is a constraint that any adequate measure of causal strength should satisfy. In this section, I want to explore whether such an argument is feasible.

The fact that Sprenger introduces GDM as a "very general adequacy constraint" (Sprenger and Hartmann, 2019, p. 160) indicates that he not only assumes it to hold for any probabilistic measure of causal strength, but also that he considers it to be rather uncontroversial. But what does really speak in favor of GDM being a generally valid adequacy constraint? I can only think of two arguments with at least some initial appeal. The first one is this: By demanding that the causal strength of a cause C on its effect E is determinable by one and the same Δ-formula in any causal scenario, GDM ensures that the determination of causal strength is rather simple. This is undoubtedly true, but simplicity does not imply adequacy. If a measure of causal strength is adequate and simple, then the simplicity is definitely a valuable pragmatic benefit. But the simplicity itself does not make it adequate. Otherwise, we could simply use $P(E|C)$ or even some constant k as a measure of causal strength of C on E, since this is even more simple than a GDM-satisfying measure and therefore, according to our simplicity-guided methodology, more adequate. I think it is obvious that this kind of methodology will lead us into trouble.

Still, one might worry that a GDM-violating measure could make the determination of causal strength values unmanageably complex. With CS_{ch}, it seems that we not only need to deduce a new Δ-formula for any new causal scenario, but the adequate Δ-formulas themselves seem to become more and more complex with increasingly complex causal scenarios. Just consider a slightly more complex variation of our example in figure 4.9, in which E has an additional generative cause A, that is probabilistically independent of all other causes of E and causally acts independently from them. The scenario is shown in figure 4.10.

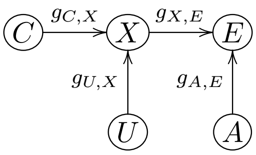


Figure 4.10: Indirect causation with externally influenced mediator and effect.

The Δ-formula that correctly determines the value of $CS_{ch}(C, E)$ in the scenario from figure 4.10 is this:[42]

$$CS_{ch}(C, E) = \frac{P(E|C) - P(E|\neg C)}{1 - P(E|\neg C) - P(X|\neg C) + P(E|\neg C) \times P(X|\neg C)} \tag{4.34}$$

The Δ-formula, that adequately determines the value of $CS_{ch}(C, E)$ in a given scenario, becomes more complex, the more external causes influence events on the active causal path from C to E. This clearly feeds the worry mentioned above: In increasingly complex causal scenarios, the determination of values of causal power becomes increasingly complex as well. But this worry is largely unfounded. Although it is true that the adequate Δ-formula for the determination of the value of $CS_{ch}(C, E)$ can get very complicated in complex causal scenarios, it also holds, that we do not necessarily need a Δ-formula to determine the value of $CS_{ch}(C, E)$. Since CS_{ch} satisfies features like Multiplicativity and Summation of Effects, we can employ compositional methods for determining values of causal power in complex causal scenarios, which means that we can determine the value of $CS_{ch}(C, E)$ as a function of the constituent causal powers involved. The constituent causal powers, on the other hand, can often be determined by much simpler Δ-formulas, like f_{ch}, because a small component of a complex causal scenario, is typically less complex. For example, we can determine the values of $CS_{ch}(C, X)$ and $CS_{ch}(X, E)$ in the scenario from figure 4.10 by $f_{ch}(C, X)$ and $f_{ch}(X, E)$, respectively. To determine the value of $CS_{ch}(C, E)$, we can then simply exploit the Multiplicativity of CS_{ch} and therefore determine $CS_{ch}(C, E) = CS_{ch}(C, X) \times CS_{ch}(X, E)$. This is how the determination of causal power values remains easily manageable even in complex causal scenarios, even for a measure that does not satisfy GDM.

The second potential argument for a compliance with GDM is this: In science and in everyday life, we often assess the causal strength of a putative cause C on a putative effect E without knowing and without caring to know whether C causally influences E directly or whether there are actually mediators between C and E. I will call this practice *measuring under mediator-*

[42] I omit the deduction of the Δ-formula in (4.34), since it works just analogously to the deduction of the Δ-formula in equation (4.29).

ignorance, or in short: *MUMI*. MUMI seems to be a common and intuitively adequate practice. A measure of causal strength that satisfies GDM, can explain why MUMI appears as such, since GDM makes sure that we can employ one and the same Δ-formula to determine the causal strength of a cause C on an effect E, no matter how the actual causal structure of the situation looks like.[43] A GDM-violating measure like CS_{ch}, on the other hand, seems to be in conflict with MUMI. How exactly the value of $CS_{ch}(C, E)$ can be determined, depends on the structure of the causal scenario, in which C and E are embedded. This clearly implies that we cannot be in ignorance of the exact underlying causal structure of a situation, when measuring values of causal power.

Although this argument seems initially appealing, I do not agree with its crucial premise, which is that MUMI is a universally adequate practice. At least it is not when it comes to intrinsic causal power. But, and this is probably why the argument has some initial appeal, I do agree that MUMI is an adequate practice under certain presuppositions and is therefore indeed a common practice in science and in everyday life. But this limited validity of MUMI is actually in accordance with employing CS_{ch} as a measure of causal power. To see why MUMI cannot be a universally adequate practice when it comes to intrinsic causal power, remember what we have shown in section 4.8.2: As soon as there is a mediator between C and E that is causally influenced by a path-external cause, the disregard of this mediator would make an adequate assessment of C's internal causal power on E impossible. If we would ignore the mediator in our assessment of the internal causal power of C on E, we would not be able to adequately balance out the floor-effect. So, when we aim for a measure of intrinsic causal power, we cannot always measure the causal power of C on E under the ignorance of whether there are any mediators between C and E. But I have also shown above, that we can use one and the same Δ-formula, namely $f_{ch}(C, E)$, to determine $CS_{ch}(C, E)$ whenever C causes E via a causal path that does not include any mediators, that are causally influenced by path-external causes.[44] So, even though CS_{ch} is incompatible with MUMI as a universally adequate practice, it can explain why MUMI is an adequate practice under the presupposition that, if there are any mediators between C and E, none of these mediators is causally influenced by a path-external cause. So, acknowledging CS_{ch} as an adequate measure of causal strength can explain why MUMI is a common and, at least under certain presuppositions, intuitively adequate practice. And, in contrast to any GDM-satisfying measure, CS_{ch} can also explain, why MUMI is not universally adequate when it comes to measuring intrinsic causal power.

4.8.5 Summary

Let me briefly summarize the results of this section. If we would simply identify Cheng's measure of causal power CS_{ch} with the formula f_{ch}, then Sprenger's results, as presented in section 4.7, would amount to a strong argument against the adequacy of CS_{ch}, because CS_{ch} would violate two highly intuitive features: Multiplicativity and Seperability of Effects. But I have argued

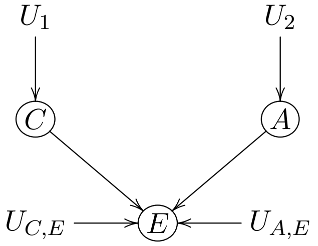
[43] Sprenger puts forward this argument, when he writes: "Notably, causal strength is blind to the presence of multiple paths leading from C to E, and to the number of mediators between C and E" (Sprenger and Hartmann, 2019, p. 160).

[44] Additionally, I will later show that $f_{ch}(C, E)$ can determine $CS_{ch}(C, E)$, if there are multiple causal paths from C to E, which do not include any mediators that are causally influenced by path-external causes.

that a general identification of CS_{ch} with f_{ch} is wrong. Cheng (1997) shows that the formula f_{ch} can indeed determine values of CS_{ch} in a very simple kind of causal scenario. But this does not mean that f_{ch} can determine values of CS_{ch} in more complex causal scenarios. I have amended Cheng's Power PC theory by, what I consider to be, highly intuitive axioms about the concept of causal production, to enable its application to certain complex causal scenarios. This expanded Power PC theory yields a measure CS_{ch} of intrinsic causal power that is not generally equatable with f_{ch} and that does not conform to GDM. I have argued that this is reasonable for a measure of intrinsic causal power. And I have also shown that CS_{ch} does indeed satisfy Multiplicativity and Seperability of Effects.

4.9 Causal Power, Causal Models, and Causal Attribution

Our explication and discussion of CS_{ch}, and of causal strength in general, has so far not included the framework of probabilistic SEMs. We have therefore not yet made clear, where and how values of causal power feature in probabilistic causal models. To do so, it is helpful to consider a simple example. We will use an SEM-representation of a Cheng-scenario. As pointed out by Glymour (1998), this causal scenario is perfectly described by an SEM, that contains the bivalent endogenous variables C, A,[45] E, and the bivalent error-terms $U_{C,E}$ and $U_{A,E}$.[46] This kind of causal model is also known as a noisy *or*-gate and is shown in figure 4.11.



- $C := U_1$
- $A := U_2$
- $E := (C \wedge U_{C,E}) \vee (A \wedge U_{A,E})$

Figure 4.11: SEM-representation of the Cheng-scenario: A noisy or-gate.

In chapter 2, I have proposed an interpretation of the error-terms in probabilistic SEMs, that, after the discussion of Cheng's Power PC theory, should sound particularly familiar. Take, for example, the error-term $U_{C,E}$. I have argued that '$U_{C,E} = 1$' should be understood as representing that, if $C = 1$ is the case, $C = 1$ causally produces $E = 1$. Accordingly, $\mathcal{P}(U_{C,E} = 1)$ amounts to the probability of $C = 1$ producing $E = 1$, given that $C = 1$ is the case, that is: $\mathcal{P}(C = 1 \text{ produces } E = 1|C = 1)$. But this is just Cheng's characterization of generative causal

[45] The variable A now represents the background causes of E. I do not use the variable U anymore, to prevent a confusion with error-terms.

[46] With the two-valued variables we again presuppose that C and A only have unambiguous contrasts $\neg C$ and $\neg A$. But notice that this presupposition is only a simplification. Multi-valued variables do not pose a problem for our considerations here. It would only demand the introduction of more error-terms.

power in axiom C1. This indicates that in a probabilistic SEM that contains a cause candidate $C = 1$ and its potential effect $E = 1$, the intrinsic causal power $CS_{ch}(C = 1, E = 1)$ of $C = 1$ on $E = 1$ will, at least under certain conditions which I will soon specify, turn up as the probability of the error-term $U_{C,E}$ taking on the value 1: $CS_{ch}(C = 1, E = 1) = \mathcal{P}(U_{C,E} = 1)$.

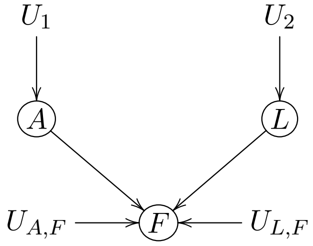
In chapters 2 and 3, I have put forward the guidelines PROBAC and PROSAC. In both guidelines, the first three steps describe the construction of a probabilistic SEM that is supposed to represent a token scenario. The first step requires the construction of a preliminary probability distribution $\mathcal{P}_{pre}$ over the exogenous variables, including the error-terms, that is, ideally, guided by information obtained from statistical data:

(1) Create a preliminary probability distribution $\mathcal{P}_{pre}$ such that: (a) $\mathcal{P}_{pre}$ ascribes probabilities to the background variables in $\mathcal{M}$ that are consistent with the current knowledge about the token scenario and (b) if $U_{X,Y}$ is an error-term that represents, whether $X = x$ causally produces $Y = y$, given $X = x$, then $\mathcal{P}_{pre}(U_{X,Y} = 1)$ should be equal to the statistically determined relative frequency with which $X = x$-type events, if present, successfully produce $Y = y$-type events.

With the Power PC theory and the measure $CS_{ch}(X = x, Y = y)$, we have now a clear explication of how the value of $\mathcal{P}(X = x \text{ produces } Y = y|X = x)$ can be determined. We can therefore now reformulate the first step of PROBAC and PROSAC in the following way:

(1) Create a preliminary probability distribution $\mathcal{P}_{pre}$ such that: (a) $\mathcal{P}_{pre}$ ascribes probabilities to the background variables in $\mathcal{M}$ that are consistent with the current knowledge about the token scenario and (b) if $U_{X,Y}$ is an error-term that represents, whether $X = x$ causally produces $Y = y$, given $X = x$, then $\mathcal{P}_{pre}(U_{X,Y} = 1) = CS_{ch}(X = x, Y = y)$.

Remember, though, that probabilistic SEMs, that represent token scenarios, can also incorporate knowledge that goes beyond pure statistical data. Just consider the probabilistic forest fire scenario from section 2.4.4, which is represented by the causal model $\mathcal{M}^F$ as shown in figure 4.12.



- $A := U_1$
- $L := U_2$
- $F := (A \wedge U_{A,F}) \vee (L \wedge U_{L,F})$

Figure 4.12: $\mathcal{M}^F$ - the probabilistic forest fire scenario.

In section 2.4.4 we assumed that the SEM represents a token scenario, about which we already know that $A = 1$ and $L = 1$ happened. We also assumed to know that the causal power of $A = 1$ on $F = 1$ is 0.7 and the causal power of $L = 1$ on $F = 1$ is also 0.7. Imagine first, that this is all we know about the scenario. Following PROBAC will then yield an SEM $(\mathcal{M}^F, \mathcal{P})$ with a probability distribution $\mathcal{P}$ that assigns the following values:

- $\mathcal{P}(U_1 = 1) = 1$
- $\mathcal{P}(U_2 = 1) = 1$
- $\mathcal{P}(U_{A,F} = 1) = 0.7$
- $\mathcal{P}(U_{L,F} = 1) = 0.7$
- $\mathcal{P}(F = 1) = 0.91$

In this probabilistic causal model $(\mathcal{M}^F, \mathcal{P})$, the probabilities of the error-terms taking on the value 1 do match the corresponding values of causal power: $\mathcal{P}(U_{A,F} = 1) = CS_{ch}(A = 1, F = 1) = 0.7$, $\mathcal{P}(U_{L,F} = 1) = CS_{ch}(L = 1, F = 1) = 0.7$. But now imagine that we get some trumping evidence and come to learn that $F = 1$ actually happened. For incorporating this information into the probabilistic SEM, we have to conditionalize $\mathcal{P}$ on $F = 1$, which gives us the updated distribution $\mathcal{P}'(\cdot) = \mathcal{P}(\cdot|F = 1)$ that, as shown in section 2.4.4, assigns the following values:

- $\mathcal{P}'(U_1 = 1) = 1$
- $\mathcal{P}'(U_2 = 1) = 1$
- $\mathcal{P}'(U_{A,F} = 1) = 0.769$
- $\mathcal{P}'(U_{L,F} = 1) = 0.769$
- $\mathcal{P}'(F = 1) = 1$

In the probabilistic SEM $(\mathcal{M}^F, \mathcal{P}')$ the probabilities of the error-terms taking on the value 1 do not match the corresponding causal power values anymore, since these are still 0.7 in both cases. As soon as we have any trumping evidence about whether the production in question did take place in the given situation, that is, any evidence, that influences our belief about whether the causal production happened and that does not come from the statistically gained data, then the probability of the error-term taking on the value 1 does not match the corresponding causal power anymore. It is therefore very well possible that in a probabilistic SEM $(\mathcal{M}, \mathcal{P})$, in which $U_{X,Y}$ is an error-term that represents whether $X = x$ causally produces $Y = y$, given $X = x$, the value of $\mathcal{P}(U_{X,Y} = 1)$ is not equal to the causal power $CS_{ch}(X = x, Y = y)$.The value of $\mathcal{P}(U_{X,Y} = 1)$ is only equal to the causal power $CS_{ch}(X = x, Y = y)$, if the probability distribution $\mathcal{P}$ has not incorporated any trumping evidence about whether the causal production of $Y = y$ by $X = x$ actually happened.[47]

[47] This also illustrates how Cheng's axiom C1 would need to be refined, if we would allow an epistemic interpretation of the probability distribution P in C1: Instead of $CS_{ch}(C, E) = P(C \text{ produces } E|C)$, it would need to say that $CS_{ch}(C, E) = P(C \text{ produces } E|C \wedge \neg T)$, where $\neg T$ represents the absence of any trumping evidence about C's production of E.

In our example, the values of $\mathcal{P}'(U_{A,F} = 1)$ and $\mathcal{P}'(U_{L,F} = 1)$ do still provide us with some useful causal information, even though they do not represent the respective causal power values anymore. Instead, they provide us with the values of what Cheng and Novick (2005) call *causal attribution*. The measure of causal attribution gives a quantitative answer to the following question: Given that the cause candidate C and the effect candidate E have actually occured, how probable is it that C has actually causally produced E? The answer lies in the value of $P(C \text{ produces } E|C, E)$. Cheng and Novick (2005) deduce the following formula for determining the causal attribution of a cause candidate C to an event E:

$$P(C \text{ produces } E|C, E) = \frac{CS_{ch}(C, E)}{P(E|C)} \tag{4.35}$$

The framework of probabilistic SEMs enables another route for determining the causal attribution of a cause candidate $C = 1$ relative to an event $E = 1$. Let $(\mathcal{M}, \mathcal{P})$ be a probabilistic SEM, in which $U_{C,E}$ is an error-term that represents whether $C = 1$ causally produces $E = 1$, given $C = 1$, and $\mathcal{P}(U_{C,E} = 1) = CS_{ch}(C = 1, E = 1) = \mathcal{P}(C = 1 \text{ produces } E = 1|C = 1)$. If we conditionalize $\mathcal{P}$ on $E = 1$ to get the updated distributiton $\mathcal{P}'(\cdot) = \mathcal{P}(\cdot|E = 1)$, then the value of $\mathcal{P}'(U_{C,E} = 1) = \mathcal{P}(U_{C,E} = 1|E = 1) = \mathcal{P}(C = 1 \text{ produces } E = 1|C = 1, E = 1)$ is the causal attribution of $C = 1$ relative to $E = 1$.[48]

4.10 Direct and Total Generative Causal Power

Glymour (1998) points to an alternative way of determining the intrinsic causal power $g_{C,E}$ of C on E in a Cheng-scenario, in which C and U are the only generative causes of E. Instead of intervening to realize C in every individual of a test-group in an RCT to measure $P(E|C)$ and intervening to realize $\neg C$ in every individual of a control-group to measure $P(E|\neg C)$ and then using these results to determine the value of $g_{C,E}$ with the formula $f_{ch}(C, E)$, we could also intervene to realize C and $\neg U$, which is the absence of every alternative cause of E besides C, in every individual of just one population. We could then simply measure the value of $P(E)$ in this population, which directly gives us the value of $g_{C,E}$. I will denote this method of determining $g_{C,E}$ by the label Gly_1. To see why Gly_1 works in a Cheng-scenario, just remember that the probability of E in such a scenario is given by:

$$P(E) = P(C) \times g_{C,E} + P(U) \times g_{U,E} - P(C) \times g_{C,E} \times P(U) \times g_{U,E} \tag{4.36}$$

It is now easy to see, that:

$$P(E|C \wedge \neg U) = g_{C,E} \tag{4.37}$$

Gly_1 is indeed a very simple method for determining the causal power of a cause-candidate C on its effect E. It is also our best chance for determining the causal power of C on E, if all possible alternatives C' to C have a causal power on E themselves.[49] Notice, though, that Gly_1

[48] See (Stephan and Waldmann, 2018), (Stephan et al., 2020), (Stephan and Waldmann, 2022), (Stephan et al., 2018) for further refinements of how to measure causal attribution.

[49] See section 4.6, where this problem is briefly discussed.

is often impracticable. The background causes U of the effect E are often not completely known. Intervening on them to “turn them off” is then clearly practically impossible.[50]

Consider now a scenario of indirect causation as shown in figure 4.13.

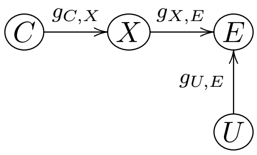


Figure 4.13: Indirect single path production.

As Glymour (1998) points out, it seems that we cannot determine the causal power $g_{C,E}$ of C on E by measuring the probability of E in a population in which we have intervened to set the relative frequency of C to 1 and the relative frequency of all other causes of E besides C to 0, since this would mean to set the relative frequency of X to 0. This would block the causal influence of C on E. As a result we would get: $P(E|C \wedge \neg X \wedge \neg U) = 0$, no matter what the value of $g_{C,E}$ is. So, Gly_1 seems to lead us astray in the given scenario. Glymour therefore proposes yet another method for determining the causal power of C on E, which I will denote by the label Gly_2: We again use only one population in which we intervene to set the relative frequency of C to 1 and the relative frequencies of all other causes of E that are not effects of C on 0. The causes of E that are effects of C are not intervened upon. Remember that the probability of E in the scenario of indirect causation is given by:

$$P(E) = P(C) \times g_{C,X} \times g_{X,E} + P(U) \times g_{U,E} - P(C) \times g_{C,X} \times g_{X,E} \times P(U) \times g_{U,E} \tag{4.38}$$

By condioning on C and $\neg U$ we get:

$$P(E|C, \neg U) = g_{C,X} \times g_{X,E} \tag{4.39}$$

With axiom C5 we have:

$$P(E|C, \neg U) = g_{C,X} \times g_{X,E} = g_{C,E} \tag{4.40}$$

So far, we have seen that each of the two methods, Gly_1 and Gly_2, can be used in certain causal scenarios to determine the value of one and the same kind of causal power, namely CS_{ch}. But this is not all to the story. While both methods, Gly_1 and Gly_2, yield the same result in the first example of direct causation, they yield different results in the second example of indirect causation. But, as Glymour points out, this should not be seen as evidence that in certain causal scenarios only one of the measures is wrong and the other one is right. Instead, the methods Gly_1 and Gly_2 point to a disambiguation of causal power: Each method measures its own kind of causal power, that only in some causal scenarios happen to coincide. As a further illustration consider an example from Glymour (1998, p. 52) that is represented by the graph in figure 4.14.

[50] A Δ-formula, like f_{ch} can deal with such situations, because it considers the causal influences of background causes on E and offsets them accordingly.

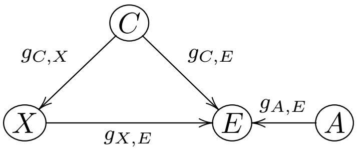


Figure 4.14: Multiple path production.

In this scenario, the probability of E is given by:

$$\begin{aligned} P(E) = {} & P(X) \times g_{X,E} + P(C) \times g_{C,E} + P(A) \times g_{A,E} - \\ & - [P(X) \times g_{X,E} \times P(C) \times g_{C,E} + P(A) \times g_{A,E} \times P(C) \times g_{C,E} + \\ & + P(X) \times g_{X,E} \times P(A) \times g_{A,E} - P(X) \times g_{X,E} \times P(C) \times g_{C,E} \times P(A) \times g_{A,E}] \end{aligned} \tag{4.41}$$

When applying Gly_1, we have to determine $P(E|C \wedge \neg X \wedge \neg A)$, which gives us:

$$P(E|C \wedge \neg X \wedge \neg A) = g_{C,E} \tag{4.42}$$

This is the probability of C producing E on the direct path, given that C is the case. Following Glymour (1998), I will denote this kind of causal power *direct causal power*. When applying Gly_2, we have to determine $P(E|C \wedge \neg A)$, which gives us:

$$P(E|C \wedge \neg A) = P(X) \times g_{X,E} + g_{C,E} - P(X) \times g_{X,E} \times g_{C,E} \tag{4.43}$$

Since $P(X) = P(C) \times g_{C,X}$ and $P(X|C \wedge \neg A) = g_{C,X}$, we have:

$$P(E|C \wedge \neg A) = g_{C,X} \times g_{X,E} + g_{C,E} - g_{C,X} \times g_{X,E} \times g_{C,E} \tag{4.44}$$

This is the probability of C producing E either on the direct or on the indirect path or on both, given that C is the case. Following Glymour (1998), I will denote this kind of causal power *total causal power*.

What does this differentiation of two kinds of causal power mean for our measure CS_{ch}? Clearly, it can only measure one of them, when both diverge. But which one is it? To find out, we should apply CS_{ch} to a scenario with multiple causal paths from C to E. But when trying to do so, another limitation of Cheng's Power PC theory reveals itself. We have not made explicit so far what it means that C produces E, if there are several different paths via which C is able to produce E. In such a situation it seems as if the concept of causal production itself is ambiguous. C can succeed in producing E on one path, while at the same time fail to produce E on a different path. One single cause C can also overdetermine E, when it succeeds in producing E on more than one path. So, what do we mean by *the* production of E by C, when there can be several productions of E by C? How can we determine the probability of C producing E, given C, when the event of C producing E is ambiguous? We need to expand Cheng's Power PC theory by yet another axiom about causal production to settle these questions. To formulate this axiom, it is helpful to first define the following concept:

Active Path of Production. *A sequence of events* $(X_1, ..., X_n)$ *is called an 'active path of production' if and only if* X_i *produces* X_{i+1} *for every* $1 \leq i \leq n-1$.

Now, here is my proposal for axiom C8:

C8 C produces E if and only if there is at least one active path of production from C to E.

As a consequence of C8, the value of $CS_{ch}(C, E) = p(C \text{ produces } E|C)$ is given by the probability that at least one path from C to E is active, given that C is the case. So, $CS_{ch}(C, E)$ always measures the total causal power of C on E, just like Gly_2, and not the direct causal power of C on E, like Gly_1, even though in some scenarios, the values of both kinds of causal power can coincide.[51] In addition to the direct and the total causal power, we clearly also have path-specific causal powers. For any specific path from C to E, the path-specific causal power is given by the probability that the given path is an active path of production, given that C is the case.

4.11 Intrinsic Causal Power vs. Contextual Causal Influence

So far, I have pretended that causal production is an activity that singular events always do on their own. But this is a major simplification. Typically, it is a combination of many factors or events that interact to causally produce an effect. We have already considered such situations. In our conjunctive forest fire scenario, for example, where only the conjunction of two singular events can causally produce a forest fire. But when the conjunctive event $C \wedge B$ has a certain generative causal power on E, while the component events C and B are not able to produce E on their own, what is the causal relation that C (or B) itself has to E? To say that C produces E, whenever $C \wedge B$ produces E is misleading, since this falsely suggests that C causally produces E on its own and that E is therefore causally overproduced by $C \wedge B$. But still, C somehow participates in the production of E. To acknowledge this participation we have to introduce a new theoretical concept, which I will call *co-production.* The following axiom describes the relation between co-production and production:

C9 If the (conjunctive) event $\bigwedge_{i \in \{1,...,n\}} A_i$ with $n \geq 2$ produces E, then for all $i \in \{1, ..., n\}$: A_i co-produces E.

The condition that $n \geq 2$ ensures that production does not imply co-production. It would be odd to say that whenever C produces E all on its own, it also co-produces E. It is possible, though, that C produces E all on its own and it additionally co-produces E by participating in another complex cause of E.[52]

In line with Cheng's axiom C1, which equates the causal power of C on E with $P(C \text{ produces } E|C)$, we will say that only those causes that can actually produce E, may it be basic or

[51] I consider this to be a consistent generalization of $CS_{ch}(C, E)$, because in the example of simple indirect causation, where Gly_1 and Gly_2 diverge, $CS_{ch}(C, E)$ can be applied even without axiom C8, since there is only one path from C to E. In that example $CS_{ch}(C, E)$ already coincides with Gly_2 and therefore diverges from the result of Gly_1.

[52] Cheng and Novick (2004) take account of this fact and, in the style of Cheng's original Power PC Theory, they assess how the causal power of two causes can be determined by a Δ-formula, if each of the causes not only exerts a causal power on the effect on their own but also an interactive causal power as a conjunctive event.

conjunctive events, do have a causal power on E. Events that can only co-produce E do not have any causal power on E on their own. But we can introduce an analogous type of causal strength for co-producing causes, whose value is given by: $P(C \text{ co-produces } E|C)$. The value of $P(C \text{ co-produces } E|C)$ is equal to the probability that C is a conjunct of at least one complex event that produces E. Consider a causal scenario, in which there is only one complex event $A \wedge C$ of which C is a conjunct and which is able to produce E. Let us denote the causal power of $A \wedge C$ on E by $g_{AC,E}$.

$$P(C \text{ co-produces } E|C) = P((C \wedge A) \text{ produces } E|C) \tag{4.45}$$

But according to axiom 2, it follows from '$(C \wedge A)$ produces E' that $C \wedge A$ is the case. We therefore have:

$$\begin{aligned} P(C \text{ co-produces } E|C) &= P((C \wedge A) \text{ produces } E|C) \\ &= P((C \wedge A) \wedge (C \wedge A) \text{ produces } E|C) \\ &= P(C \wedge A|C) \times P((C \wedge A) \text{ produces } E|C \wedge (C \wedge A)) \\ &= P(A|C) \times P((C \wedge A) \text{ produces } E|C \wedge A) \\ &= P(A|C) \times g_{AC,E} \end{aligned} \tag{4.46}$$

This already illustrates that the value of $P(C \text{ co-produces } E|C)$ is highly contextual, since it depends on the probabilities of C's co-producers. So, unlike $CS_{ch}(C, E)$, it is not a measure of *intrinsic* causal power. This is why I will denote this type of causal strength as the *contextual causal influence* of C on E, formally: $CI(C, E)$.[53]

Notice, though, that the contextual causal influence of a co-producer can, at least in a restricted range, sometimes be just as invariant under context changes and therefore just as useful for causal predictions as an intrinsic causal power. Imagine a causal producer $A \wedge C$ of E with an intrinsic causal power $g_{AC,E}$ and with A and C being probabilistically independent. We then have: $CI(C, E) = P(A) \times g_{AC,E}$. Now, if the probability $P(A)$ of A remains constant for all contexts of interest, then the value of $CI(C, E)$ is just as invariant across those contexts as the intrinsic causal power $g_{AC,E}$ of the complete producer $A \wedge C$.[54]

4.12 Summary

In this chapter, I have introduced Cheng's Power PC theory and the measure CS_{ch} that results from it. I have argued that CS_{ch} is the most promising candidate for a measure of intrinsic

[53] Cheng (2000) calls it 'contextual power'. But I consider this label somewhat misleading, since the word 'power' still suggests an intrinsic property of the cause candidate.

[54] In the framework of causal models, it is a common practice that certain co-producers A of a complete producer $A \wedge C$ of E are neglected and not explicitly represented as endogenous variables in the model, if it is assumed that $P(A)$ remains constant over all relevant contexts and if a manipulation of factors in A is either impossible or of no interest. In such cases, the factors in A enter the causal model anonymously in form of an error-term. This is why, even in deterministic systems where all genuine intrinsic causal powers have only trivial values, we often work with causes that have non-trivial 'causal powers'. These causal powers are actually just contextual causal influences that are invariant across all relevant contexts.

causal power that is supposed to remain invariant across different contexts, which makes it an indispensable tool for causal predictions, causal explanations, decision making, and targeted manipulations. I have argued that Sprenger's arguments against Cheng's measure are misguided, since they falsely identify CS_{ch} with a formula that can only occasionally determine the values of CS_{ch}. As a response, I have expanded Cheng's Power PC theory to enable the determination of intrinsic causal powers in more complex causal scenarios as well. In the last three sections of this chapter, I have differentiated the concept of intrinsic causal power from causal attribution and the contextual causal influence of a co-producer. I have dicussed the differentiation of causal power into direct, total, and path-specific causal power and I have illustrated how causal powers feature in probabilistic structural equation models.

Chapter 5

Preventive Power

5.1 Prevention as Production of the Complement

Causal production has an adversary. Just like there are events that causally produce other events, there are events that prevent other events. Turning on the heating on a cold winter day prevents a tenant from freezing. Drinking prevents dehydration. Taking a certain medicine prevents high blood pressure. Prevention is as omnipresent in our causal perception of the world as is causal production. But what exactly is it? A popular way to define prevention is to identify it with a certain form of causal production:

Prevention as Production of the Complement (PPC). *C prevents E if and only if C produces $\neg E$.*

According to Nancy Cartwright, this identification "follows the usual conventions" (Cartwright, 1989, p. 99). Reducing prevention to causal production in this way certainly has an intuitive appeal. Consider as an example Suzy and Billy who, after running out of empty bottles, start a tug war. They both pull as hard as they can. But shortly before Billy looses his grip a strong gust of wind blows in Billy's direction. Suzy can't defy both Billy and the wind and looses. Clearly, the wind gust prevented Suzy's win, just like it produced her defeat. It seems natural to say that both formulations are equivalent. In the spirit of PPC Fitelson and Hitchcock (2011), as well as Sprenger (2018), equate the absolute value of preventive strength $PS(C, E)$ of a preventive cause C of E with the absolute value of generative causal strength $CS(C, \neg E)$ of C on $\neg E$:

Preventive Power as Generative Power on the Complement. $\mid PS(C, E) \mid = \mid CS(C, \neg E) \mid$.

So, if PPC is correct, then all we need for assessing the preventive power of a preventive cause is a measure of generative causal power.

5.2 The Intuition of Difference

But is prevention really equatable with production of the complement? There are reasons to doubt it.[1] Consider the following example by Dowe (2001): A child runs onto a street without paying attention to the traffic. A car would hit the child, if it would continue its path. But, luckily, the father grabs the child just in time and thereby prevents the accident. Dowe (2001) presents this example to illustrate, what he calls, the *intuition of difference*: While it seems intuitively clear, that the father, by grabbing the child, prevented the accident, it seems intuitively dubious to say that the father, by grabbing the child, causally produced the complement of the accident, namely the non-occurrence of the accident. So, intuitively, we have an instance of prevention without production of the complement.

Dowe (2001) suggests that the intuition of difference in the given example is due to the fact that the non-occurence of the accident does not really seem to be a genuine event. It rather seems to be a highly disjunctive abstraction that can be realized in arbitrarily many ways and it is not clear what kind of objects, activities and properties belong to it. Intuitively, only genuine events can serve as causal relata. But, according to Dowe (2001), the complement of a genuine event is often no genuine event itself, but rather, what he calls, a "negative event" (Dowe, 2001, p. 216). Dowe (2001) supposes that the intuition of difference is always rooted in such negative events and the fact that negative events cannot be relata of causal production. So, as soon as an event prevents another event, whose complement is a negative event, we have a counterexample to PPC, since we have prevention without production of the complement.

I am not convinced by Dowe's explanation of the intuition of difference. In the given example, it rather appears to me as a rhetorical ploy. To name the complement of an accident "non-occurence of the accident" suggests that nothing happens, which, of course, cannot be causally produced, because there is nothing to produce. A slight reformulation changes the impression, though. The accident in question basically amounts to the occurence of drastic deformations to the car as well as injuries to the child at a certain time t. The complement to this would be that the car as well as the child display just or nearly the same physical condition at time t as they did shortly before t. Now, the problem is not that the physical conditions of the car and of the child at time t cannot be causally produced in the first place. It is rather that the father's grabbing of the child had no part in causally producing them. Further, the fact that the description of an event is abstract or highly disjunctive does not make it an event, that cannot be causally produced. An accident is itself highly disjunctive and can be realized in arbitrarily many ways. But it makes sense to say that an event can causally produce an accident, if it causally produces a certain concrete manifestation of, what we somewhat vaguely call, an accident. I think the same holds for the complement. An event can causally produce a non-accident, if it produces a concrete event that is a manifestation of the abstract description. We have to keep in mind, that we are not talking about any accident or non-accident, but about an accident or non-accident that happens at a specific place, a specific time and that involves a specific child. This restricts drastically how such a manifestation might look like.

[1] The "if and only if" in PPC does not have to be understood as an identification of prevention with production of the complement. It could also be read as saying that they always co-occur. But, as the following examples show, even this weaker reading is very disputable.

While I agree with Dowe that there is an intuition of difference between prevention and production of the complement, I am not convinced that this intuition is always due to the fact that the complement of a given genuine event is a negative event, that cannot be causally produced. And even if there really are some instances, where this is the case, it is definitely not always the case. There are cases, where the intuition of difference occurs, even though the complement of the prevented (genuine) event is a clear-cut specimen of a genuine event. Imagine, for example, three balls B_1, B_2 and B_3. None of the three balls have been in contact before, but, due to some prior collisions with other balls, B_1 moves with a velocity smaller than v. B_2 is on a collision course with B_1 and since B_2 has a high momentum, the collision would cause B_1 to have a velocity higher than v at time t. But B_3 collides first with B_2 and thereby prevents its collision with B_1. Intuitively, B_3's collision with B_2 prevented B_1 from having a velocity higher than v at t. But B_3's collision with B_2 did, intuitively, not produce or co-produce B_1 having a velocity lower than v at t. Instead, this event was produced by prior collisions with other balls. We thus have a clear example of an intuition of difference, even though the complement of the prevented event is itself undoubtedly a genuine event.

It might come across as an overreaction to topple a well established theory like PPC just because two individuals have the intuition that it is wrong. Fortunately, Walsh and Sloman (2011) has made sure, that the intuition of difference has obtained some experimentally validated backup. The authors have confronted fifty-seven research participants with the following scenario:

> "There is a coin standing on its edge at the end of the table. It is unstable and it is about to fall over and land on heads. Frank and Jane are standing close by with marbles. While they are there someone else rolls a marble toward the coin. The roll is perfectly on target and it will hit the coin, knock it over and the coin will land on tails. Frank and Jane both reach out and put their hands in front of the coin. Frank happens to put his hand in front of Jane's and he catches the marble. The coin falls over and lands on heads" (Walsh and Sloman, 2011, p. 39).

The participants had to answer the following questions:[2]

> "Did Frank cause the coin to land on heads?
> Did Jane cause the coin to land on heads?
> Did Frank prevent the coin from landing on tails?
> Did Jane prevent the coin from landing on tails?"

65% of the participants considered Frank's action as a prevention of the coin landing on tails. But only 23% of the participants considered Frank's action as a cause of the coin landing on heads. The authors therefore conclude: "our results showed clearly that participants do not judge prevent to mean the same thing as to cause it not to occur" (Walsh and Sloman, 2011, p. 42). And, here again, the complement of the prevented event is a clear-cut specimen of a genuine event.

[2] See (Walsh and Sloman, 2011, p. 39).

Notice, though, that the intuition of difference does not entail the claim that there are never any preventers of some given event E that are also producers of the complement $\neg E$. This may very well be the case. The intuition of difference only denies that every preventer must be a producer of the complement. The intuition of difference tells us that prevention cannot be reduced to production of the complement and, consequently, that preventive power cannot be reduced to generative causal power on the complement.

5.3 Prevention as a Three-Place Relation

Hiddleston (2005) points to another intuitive perception of prevention that corroborates the intuition of difference. He points out that "[m]any preventers are derivative from generative causes. These are not simple causes of $\neg E$, but preventers *of specific causes of* E. For example, an antidote prevents death, but in the first instance it prevents a specific poison from causing death. An antidote need have no influence on death absent the specific poison to which it is an antidote" (Hiddleston, 2005, p. 41. *Emphasis in original.*) In such cases, prevention is not perceived as a two-place relation between a preventer A and a prevented effect E. It is instead seen as a three-place relation that holds between a preventer A, a generative cause C and an effect E: A prevents C from producing E. This perception of prevention makes a simple identification with the two-place relation of *production of the complement* (A produces $\neg E$) even less attractive.

Hiddleston's (2005) formulation, that "[m]any preventers are derivative from generative causes", suggests that there are still preventive causes that can prevent an effect E per se and not only by preventing a certain generative cause from producing E. This would mean that there are two structurally different types of prevention: A two-place relation of prevention (A prevents E per se) and a three-place relation of prevention (A prevents C from producing E). I think that this assumption would amount to an unnecessary inflation of the conceptual tools for our causal understanding of the world. Clearly, there are preventive causes A that can prevent several different causes from producing E. This means that A stands in a three-place prevention relationship with E and several distinct generative causes of E: A prevents C from producing E, A prevents B from producing E, and so on. So, if there really is a preventive cause A that can prevent an event E "per se", this ostensible two-place relation of prevention can be understood as a set of three-place relations of prevention: For every generative cause C of E, A can prevent C from producing E. This is why, in the following, I will treat prevention generally as a three-place relation between a preventive cause A, a prevented event E and a generative cause C of E.

5.4 An Axiomatic Approach to Preventive Power

Having reached the conclusion that prevention cannot be identified with production of the complement and that, consequently, preventive power on E cannot be equated with generative causal power on $\neg E$, we now face the question: How else can we measure preventive power? So far, the Power PC theory, as I have presented and expanded it in the previous chapter, is based only on the theoretical concepts of causal production and generative causal power. It is therefore not yet equipped to handle cases of prevention. In the present section, I aim to change that. Unlike the PPC-approach, I will not try to somehow reduce the concept of prevention to a certain form of production. Instead, I will treat the concept of prevention, in analogy to the concept of production, as a basic, theoretical, and unobservable concept. Just like the concept of production, the concept of prevention will receive its meaning only from the axioms that we incorporate into the Power PC theory. I will augment the Power PC theory with additional axioms about the concepts of prevention and preventive power, that will ultimately allow me to explicate, how preventive causal power can be measured.

Let me start with the following natural assumption about the concept of prevention:

C10 If A prevents C from producing E, then A is the case, C is the case, and C does not produce E.

C10 clearly treats prevention as a three-place relation, as argued in the previous section. Notice that A's prevention of C's production of E does not imply that E is not the case. E may very well happen, although A successfully prevents C from producing E, since a different cause than C might still produce E. Causal prevention, just like causal production, is nonetheless a success term. But the success is that C, despite being the case, does not produce E, and not that E does not occur.

To illustrate the motivation for the next axiom, consider the following example: There is a virus C that, when it has entered the human body, produces a certain desease E with probability 0.1. C is the only possible cause of E. But there is a vaccine A that prevents C from producing E with absolute certainty, which means: there is no case, in which A fails to prevent C's production of E. Now, consider a morally problematic trial with 100 persons that are all injected with the virus and the vaccine. The question is: In how many cases did A actually prevent C from producing E? It would be quite strange to say that this happened in all 100 cases, because even without the vaccine the virus would have produced the desease only in 10 percent of the 100 cases. It seems very unintuitive to claim that the vaccine prevented the virus from producing the desease in cases, in which the virus would not have produced the desease in the first place. So, the intuitively correct answer to the question above is: A prevented C from producing E in 10 cases. This intuition is captured by the following axiom:

C11 Given that there is no other preventer of C's production of E, A only prevents C from producing E, if C would have produced E, if A would not have been the case.

C11 explains why preposterous claims, like 'Dancing ridiculously prevents pop music from producing cancer', seem so preposterous. The reason is simply that pop music would not produce cancer, even if one would not dance ridiculously. Axiom C11 basically ensures that A prevents C's production of E only if A is a de facto difference maker for C's production of E.[3]

[3] A must only be a *de facto* difference maker for C's production of E, instead of a full-fledged difference maker, because there might be a "back-up" preventer B of C's production of E. In such a case, A might very well prevent C's production of E, even though C would still not have produced E, if A would not have been the case, simply because B would then have prevented C's production of E.

With these two assumptions about the concept of prevention at hand, we can now come to the concept of preventive power. In analogy to Cheng's characterization of generative causal power in C1, which is based on the concept of causal production, Hiddleston (2005) proposes a similar characterization of preventive power, that is based on the concept of prevention:[4,5]

Preventive Power (Hiddleston, 2005). *The (C-specific) preventive power $h_{A(C),E}$ of A to prevent C from producing E is given by the probability that A prevents C from producing E, given A and C: $P(A$ prevents C from producing $E|A \wedge C)$.*

At a first glance, this characterization seems reasonable. It clearly treats prevention as a three-place relation, as Hiddleston (2005) himself has argued for, and it clearly mirrors the basic idea of Cheng's definition of generative causal power. While the generative causal power determines the probability of C's production of an effect E, given that C is the case, the preventive power determines the probability of A's prevention of C's production of E, given that $A \wedge C$ is the case. But a closer look reveals, that Hiddleston's characterization of preventive power is problematic. Consider again the virus-example. In our morally reprehensible, though fictitious, experiment we have injected 100 people with the virus C and with the vaccine A. We presupposed that C's causal power to produce a certain desease (E) in the human body is 0.1. But the vaccine prevents C from producing E with absolute certainty. So, intuitively, it clearly has maximal preventive power. The value of $h_{A(C),E}$ should therefore be 1. Now, what is the value of $P(A$ prevents C from producing $E|A \wedge C)$? Because of C11, A prevents C from producing E in 10% of all cases, since C would produce E in 10% of all cases and A prevents this production every time. But $A \wedge C$ is the case in 100% of all cases. This gives us:

$$P(A \text{ prevents } C \text{ from producing } E|A \wedge C) = \frac{0.1 \times 1}{1} = 0.1 \tag{5.1}$$

This differs widely from the intuitively correct value of 1. The example also illustrates that, according to Hiddleston's characterization of preventive power, the power of A to prevent C from producing E is highly dependent on the generative causal power of C on E. Any change in the generative causal power of C on E leads to a corresponding change in the power of A to prevent C from producing E. Preventive power would accordingly not be an intrinsic feature of the preventive cause.

There are two ways out of this problem. First, one could stick to Hiddleston's characterization of preventive power and reject axiom C11. But this would be a very high price to pay. Without C11 an event A could prevent another event C from producing an effect E, even if C would not have produced E in the first place. So, nothing stops us from declaring that dancing ridiculously prevents pop music from producing cancer. C11 is our best weapon against wild assertions like that. This leaves us only with the second option: We have to acknowledge that Hiddleston's characterization of preventive power is not accurate and we have to come up with another proposal. So, here it is:

[4] See (Hiddleston, 2005, p. 42). I have slightly adapted the notation.

[5] Although Cheng (1997) deals with prevention and preventive power in her Power PC theory, she does not explicitly formulate a characterization of preventive power that stands in analogy to axiom C1. She only deduces a Δ-formula that is supposed to measure the preventive power of a preventive cause in a very simple causal scenario. I will discuss this Δ-formula in a later section and I will show how it corresponds to the conception of preventive power that I will develop here.

C12 The (C-specific) preventive power $h_{A(C),E}$ of A to prevent C from producing E is the probability that A prevents C from producing E, given A and C and the fact that C would have produced E, if A would not have been the case, given that no other preventer of C's production of E is present. More formally: $h_{A(C),E} = P(A$ prevents C from producing $E|A \wedge C \wedge \gamma_{C,E} \wedge \neg\pi)$, with $\gamma_{C,E}$ standing for "C would have produced E, if A would not have been the case" and $\neg\pi$ standing for "No other C-specific preventer of E besides A is present".

Applying this characterization of preventive power to the virus-example gives us:

$$P(A \text{ prevents } C \text{ from producing } E|A \wedge C \wedge \gamma_{C,E} \wedge \neg\pi) = \frac{0.1}{0.1} = 1 \tag{5.2}$$

C12 ensures that we put the frequency of A's preventions of C's production of E not in relation to all cases in which A and C are present, but in relation to all cases, in which A and C are present and in which C would actually have produced E, if it would not have been prevented by A or any other preventer. With this change, we put the number of actual preventions in relation to the number of cases, in which a prevention is possible in the first place. It thereby also ensures that A's power to prevent C from producing E is independent from C's generative causal power on E, which means that it can be seen as an intrinsic feature of A.

With a characterization of preventive power at hand, a crucial question remains: Can we explicate, just like we did for generative causal power, an empirical procedure for measuring the preventive power of a preventive cause? The answer is yes. But before we can do so, we first have to come back to the concept of generative causal power itself. The consideration of preventers forces us to make some adjustments in our previous explications of generative causal power.

5.5 Effective Causal Influence

In chapter 4, I have argued that it is a characteristic property of generative causal power to be seen as an intrinsic capacity of its bearer which remains invariant across different contexts. But this feature of non-contextuality seems to falter as soon as we consider scenarios, in which generative causes are confronted with preventive causes. Let C be a generative cause of E with generative causal power $g_{C,E}$ and let A be a C-specific preventive cause of E with a preventive power $h_{A(C),E} > 0$. Axiom C1 tells us that $g_{C,E} = P(C$ produces $E|C)$. But since A has a non-zero probability to prevent a production of E by C, that would otherwise have succeeded, it follows that $P(C$ produces $E|C)$ is smaller when A is present than when A is not present. But this means that the causal power of C on E takes on different values, depending on the context, or more specifically, depending on the probability of A being present. So, considering the possibility of preventive causes, generative causal power, as characterized by C1, turns out to be highly contextual, that is non-intrinsic, after all.

We basically face two options now: Either we accept the contextuality of generative causal power or we adjust C1 to obtain a concept of causal strength that remains intrinsic, even when considering the possibility of preventive causes. The decision between the two options is not really a matter of right or wrong. It is rather a decision about what kind of causal concepts we aim to use and consider more useful. I have already highlighted the value of having a quantity of causal power that can be assumed to remain constant across different contexts. It is for this reason that I will go for the second option and propose the following amendment of axiom C1:[6]

C1* The generative causal power $g_{C,E}$ of C to produce E is the probability that C produces E, if C is the case and no C-specific preventers of E are present: $P(C \text{ produces } E|C \wedge \neg A)$, with A representing all C-specific preventive causes of E.

C1* restores the non-contextuality of causal power, even for contextual changes that include changes about the prevelance of preventers. But what exactly happens to a generative cause C and its causal influence on E, when a C-specific preventer of E is present. If C's generative causal power on E is intrinsic and a C-specific preventer of E does not lower C's generative causal power, then how can we explain that the probability of C producing E is lowered in scenarios, in which A is present, in comparison to scenarios, in which A is not present? The answer is that, whenever a generative cause C of E and a C-specifc preventer of E are both present, two intrinsic causal powers, a generative causal power and a preventive causal power, interfere. This interference of two separate forces leads to a compounded influence on E. In presence of a C-specific preventer, we therefore have to differentiate between the intrinsic generative causal power of C on E and the *effective causal influence* of C on E, which is a result of the interference of C's generative causal power on E with the preventive power of a C-specific preventer of E.[7] The effective causal influence of C on E can therefore be seen as the amount of C's causal power that, despite mitigating forces, actually affects E. I will write $ECI^{\{A\}}_{C,E}$ for the effective causal influence of C on E in the presence of the C-specific preventer A, which is given by $P(C \text{ produces } E|C \wedge A \wedge$ no other C-specific preventers of E are present$)$. $ECI^{\{A_1,...,A_n\}}_{C,E}$ is the effective causal influence of C on E in the presence of the C-specific preventers $A_1, ..., A_n$. Similarly, I write $ECI^{\varnothing}_{C,E}$ for the effective causal influence of C on E in the absence of all C-specific preventers, which is just the intrinsic causal power of C on E. The empirical determination of the value of $ECI^{\{A_1,...,A_n\}}_{C,E}$ works just like the empirical determination of the value of generative causal power $CS_{ch}(C, E)$ as discussed in chapter 4, with the only difference that we keep the relative frequency of $A_1, ..., A_n$ fixed at 1 in both the control- and the test-goup of an RCT.

5.6 Measuring Preventive Power

Equipped with the differentiation between intrinsic causal power and effective causal influence, we can now follow up on the question of how to measure preventive power.

5.6.1 A General Method for Measuring Preventive Power

Let C be a generative cause of E and let A be a C-specific preventer of E. $\gamma_{C,E}$ represents the counterfactual "C would have produced E, if A would not have been present" and $\neg\pi$ represents

[6] Hiddleston (2005) proposes the same amendment of Cheng's characterization of generative causal power.

[7] Here I only consider preventive causes that mitigate the causal influence of a generative cause on its effect. But the causal influence of a generative cause C of E might also be enhanced by the presence of other factors.

"Besides A, there is no other C-specific preventer present". According to C12, the C-specific preventive power of A on E is given by:

$$h_{A(C),E} = P(A \text{ prevents } C \text{ from producing } E | A \wedge C \wedge \gamma_{C,E} \wedge \neg\pi) \tag{5.3}$$

A prevents C from producing E if and only if A satisfies C10 and C11, given that no other C-specific preventer of E is present. So we get:

$$\begin{aligned} h_{A(C),E} &= P(A \text{ prevents } C \text{ from producing } E | A \wedge C \wedge \gamma_{C,E} \wedge \neg\pi) \\ &= P(A \wedge C \wedge \neg(C \text{ produces } E) \wedge \gamma_{C,E} | A \wedge C \wedge \gamma_{C,E} \wedge \neg\pi) \\ &= P(\neg(C \text{ produces } E) | A \wedge C \wedge \gamma_{C,E} \wedge \neg\pi) \\ &= 1 - P(C \text{ produces } E | A \wedge C \wedge \gamma_{C,E} \wedge \neg\pi) \\ &= 1 - \frac{P(C \text{ produces } E \wedge \gamma_{C,E} | A \wedge C \wedge \neg\pi)}{P(\gamma_{C,E} | A \wedge C \wedge \neg\pi)} \end{aligned} \tag{5.4}$$

If we presuppose that A can only prevent, but not enhance C's production of E,[8] then there is no single instance, in which C produces E while A is the case and C would not have produced E, if A would not have been the case. This means that $\gamma_{C,E}$ holds whenever C produces E. So we get:

$$\begin{aligned} h_{A(C),E} &= 1 - \frac{P(C \text{ produces } E \wedge \gamma_{C,E} | A \wedge C \wedge \neg\pi)}{P(\gamma_{C,E} | A \wedge C \wedge \neg\pi)} \\ &= 1 - \frac{P(C \text{ produces } E | A \wedge C \wedge \neg\pi)}{P(\gamma_{C,E} | A \wedge C \wedge \neg\pi)} \\ &= \frac{P(\gamma_{C,E} | A \wedge C \wedge \neg\pi) - P(C \text{ produces } E | A \wedge C \wedge \neg\pi)}{P(\gamma_{C,E} | A \wedge C \wedge \neg\pi)} \end{aligned} \tag{5.5}$$

By our definition of $\gamma_{C,E}$, we have: $P(\gamma_{C,E} | A \wedge C \wedge \neg\pi) = P(C$ would have produced E, if $\neg A$ would have been the case $| A \wedge C$ and no other C-specific preventer besides A is present). But the probability that C produces E under the supposition, that C is the case and no preventer of C's production of E is present, is nothing else than the intrinsic generative causal power of C on E. So, we have: $P(\gamma_{C,E} | A \wedge C \wedge \neg\pi) = P(C$ produces $E | C \wedge$ no C-specific preventer of E is present) $= ECI^{\varnothing}_{C,E}$. According to our definition of effective causal influence, we further have: $P(C \text{ produces } E | A \wedge C \wedge \neg\pi) = ECI^{A}_{C,E}$. This gives us:

$$h_{A(C),E} = \frac{ECI^{\varnothing}_{C,E} - ECI^{\{A\}}_{C,E}}{ECI^{\varnothing}_{C,E}} \tag{5.6}$$

Notice that, whenever a quantity decreases from a value x to a value y, the percentage decrease p from x to y is given by the following formula:

$$p = \frac{x - y}{x} \tag{5.7}$$

[8] If A is indeed an event that has both, a preventive and an enhancing influence on C's production of E, then it is impossible to measure the two kinds of influences independently, without somehow blocking one of them.

We can therefore define the following, generally applicable measure CS_{prev} of preventive power:

Measure of Preventive Power (CS_{prev}). *The C-specific preventive power $h_{A(C),E}$ of A on E can be measured by the percentage decrease from C's intrinsic generative causal power $ECI^{\varnothing}_{C,E}$ on E to C's effective causal influence $ECI^{\{A\}}_{C,E}$ on E:*

$$CS_{prev}(A(C),E) := \frac{ECI^{\varnothing}_{C,E} - ECI^{\{A\}}_{C,E}}{ECI^{\varnothing}_{C,E}} \tag{5.8}$$

In chapter 4, we have already discussed empirical methods for measuring the value of $ECI^{\varnothing}_{C,E}$. In the previous section, I have pointed out that the value of $ECI^{\{A\}}_{C,E}$ can be determined in just the same way, with the only difference, that the probability of A needs to be kept fixed at 1 in the test- and the control-group. With these empirical methods at hand, we now also have an empirical method for determining the power $h_{A(C),E}$ of A to prevent C from producing E: We first determine the intrinsic generative causal power $ECI^{\varnothing}_{C,E}$ of C on E empirically. Next, we determine the effective causal influence $ECI^{\{A\}}_{C,E}$ of C on E in the presence of A. We then obtain the value of $h_{A(C),E}$ by computing the percentage decrease from $ECI^{\varnothing}_{C,E}$ to $ECI^{\{A\}}_{C,E}$.

Notice that, whenever we know the power $h_{A(C),E}$ of A to prevent C from producing E, as well as the generative causal power $g_{C,E} = ECI^{\varnothing}_{C,E}$ of C on E, we can easily determine the effective causal influence $ECI^{\{A\}}_{C,E}$ of C on E in the presence of A and the absence of all other C-specific preventers, by the following formula:

$$ECI^{\{A\}}_{C,E} = g_{C,E} \times (1 - h_{A(C),E}) \tag{5.9}$$

So far, we have only deduced a method for determining $h_{A(C),E}$, in which we have to ensure that no other C-specific preventers of E besides A are present. But we might not always be able to ensure this. So, can we also determine the power of A to prevent C from producing E, if there is another C-specific preventer, say B, present, which already leads to a mitigated causal influence of C on E? Luckily, the answer is yes. At least as long as both preventers act independently of each other, which means that the probability of B preventing C from producing E and the probability of A preventing C from producing E are mutually independent. The value of $h_{A(C),E}$ is given by the percentage decrease of the effective causal influence of C on E that is due to A. This percentage decrease remains the same, even if there is already a background-factor B that decreases the effective causal influence of C on E. To see this, notice that if A and B influence C's production of E independently of each other, then we have:

$$ECI^{\{A,B\}}_{C,E} = g_{C,E} \times (1 - h_{B(C),E}) \times (1 - h_{A(C),E}) \tag{5.10}$$

Consequently:

$$
\begin{aligned}
\frac{ECI_{C,E}^{\{B\}} - ECI_{C,E}^{\{B,A\}}}{ECI_{C,E}^{\{B\}}} &= \frac{g_{C,E} \times (1 - h_{B(C),E}) - g_{C,E} \times (1 - h_{B(C),E}) \times (1 - h_{A(C),E})}{g_{C,E} \times (1 - h_{B(C),E})} \\
&= \frac{g_{C,E} - g_{C,E} \times (1 - h_{A(C),E})}{g_{C,E}} \\
&= \frac{ECI_{C,E}^{\{\varnothing\}} - ECI_{C,E}^{\{A\}}}{ECI_{C,E}^{\{\varnothing\}}} \\
&= h_{A(C),E}
\end{aligned}
\tag{5.11}
$$

5.6.2 Cheng's Formula for Measuring Preventive Power

In analogy to f_{ch}, which measures the causal power of a generative cause in a Cheng-scenario, Cheng (1997) deduces a Δ-formula that is supposed to measure the intrinsic power of a preventive cause in a similarly simple causal scenario. I will now briefly present how Cheng deduces this formula, to then discuss how it corresponds to our explication of preventive power as developed in the previous sections.

Cheng considers the following causal scenario: There is one generative cause C of E with the generative causal power $g_{C,E}$ and a C-specific preventive cause A of E with the preventive power $h_{A(C),E}$.[9] It is further assumed that $h_{A(C),E}$, $P(C)$, $P(A)$, and $g_{C,E}$ are mutually independent. Cheng (1997, p. 375) points out that the following formulas hold in this scenario:

$$P(E|A) = P(C) \times g_{C,E} \times (1 - h_{A(C),E}) \tag{5.12}$$

$$P(E|\neg A) = P(C) \times g_{C,E} \tag{5.13}$$

This is well in line with our explication of $h_{A(C),E}$, because the probability of E, given A, is given by the probability of C multiplied with the probability that C produces E, given C and A, which is $ECI_{C,E}^{A}$. So, we have $P(E|A) = P(C) \times ECI_{C,E}^{A}$. Since $ECI_{C,E}^{A} = g_{C,E} \times (1 - h_{A(C),E})$, we get: $P(E|A) = P(C) \times g_{C,E} \times (1 - h_{A(C),E})$. And since we assume that no other factor causally influences E besides A and C, we have $P(E|\neg A) = P(C) \times g_{C,E}$. The next step is already familiar from the deduction of f_{ch}:

$$
\begin{aligned}
P(E|\neg A) - P(E|A) &= P(C) \times g_{C,E} - P(C) \times g_{C,E} \times (1 - h_{A(C),E}) \\
&= P(C) \times g_{C,E}(1 - (1 - h_{A(C),E})) \\
&= P(C) \times g_{C,E} \times h_{A(C),E}
\end{aligned}
\tag{5.14}
$$

Solving for $h_{A(C),E}$ gives us:

[9] Cheng does not speak of a 'C-specific' preventive cause. Instead, she treats prevention as a two-place relation. But in the given scenario with only one generative cause C, this makes no difference for the determination of $h_{A(C),E}$.

$$\begin{aligned} h_{A(C),E} &= \frac{P(E|\neg A) - P(E|A)}{P(C) \times g_{C,E}} \\ &= \frac{P(E|\neg A) - P(E|A)}{P(E|\neg A)} \\ &= \frac{-\Delta P_{A,E}}{P(E|\neg A)} =: f_{ch-prev}(A,E) \end{aligned} \tag{5.15}$$

For the given causal scenario, the application of Cheng's formula $f_{ch-prev}(A,E)$ yields just the same result as applying our measure $CS_{prev}(A(C),E)$, which determines $h_{A(C),E}$ by measuring the percentage decrease from $ECI^{\varnothing}_{C,E}$ to $ECI^{\{A\}}_{C,E}$. To see this, consider that the following holds:

$$\begin{aligned} \frac{P(E|\neg A) - P(E|A)}{P(E|\neg A)} &= \frac{P(C) \times g_{C,E} - P(C) \times g_{C,E} \times (1 - h_{A(C),E})}{P(C) \times g_{C,E}} \\ &= \frac{g_{C,E} - g_{C,E} \times (1 - h_{A(C),E})}{g_{C,E}} \\ &= \frac{ECI^{\varnothing}_{C,E} - ECI^{\{A\}}_{C,E}}{ECI^{\varnothing}_{C,E}} \end{aligned} \tag{5.16}$$

But in other, more complex causal scenarios, $f_{ch-prev}(A,E)$ fails to determine the preventive power of A on E, relative to a generative cause C of E. Consider, for example, the following scenario: We have two generative causes, C and B, that influence E independently of each other. Imagine that $P(C) = 1$, $P(B) = 0.5$, $g_{C,E} = 0.5$, and $g_{B,E} = 0.8$. We further have the C-specific preventer A of E with $h_{A(C),E} = 0.6$. But A does not have any B-specific preventive power on E, so: $h_{A(B),E} = 0$. We then have:

$$\begin{aligned} P(E|\neg A) &= P(C) \times g_{C,E} + P(B) \times g_{B,E} - P(C) \times g_{C,E} \times P(B) \times g_{B,E} \\ &= 0.5 + 0.4 - 0.5 \times 0.4 \\ &= 0.7 \end{aligned} \tag{5.17}$$

$$\begin{aligned} P(E|A) &= P(C) \times g_{C,E} \times (1 - h_{A(C),E}) + P(B) \times g_{B,E} - \\ &- P(C) \times g_{C,E} \times (1 - h_{A(C),E}) \times P(B) \times g_{B,E} \\ &= 0.5 \times (1 - 0.6) + 0.4 - 0.5 \times (1 - 0.6) \times 0.4 \\ &= 0.2 + 0.4 - 0.2 \times 0.4 \\ &= 0.52 \end{aligned} \tag{5.18}$$

So, we get:

$$f_{ch-prev}(E,A) = \frac{P(E|\neg A) - P(E|A)}{P(E|\neg A)} = \frac{0.7 - 0.52}{0.7} = 0.256 \tag{5.19}$$

This shows, the value of $f_{ch-prev}(A,E)$ does not accord with A's C-specific preventive power on E, which is $h_{A(C),E} = 0.6$. This should not be surprising. $f_{ch-prev}(A,E)$ is only a function of $P(E|A)$ and $P(E|\neg A)$. In the presence of several different generative causes of E, $f_{ch-prev}(A,E)$ can therefore not differentiate between the cause-specific preventive powers that A has relative to each generative cause of E. Instead of yielding the preventive power, that A has relative to a certain generative cause of E, $f_{ch-prev}(A,E)$ provides us with the percentage decrease from the probability that any cause produces E, given $\neg A$, which is $P(E|\neg A)$, to the probability that any cause produces E, given A, which is $P(E|A)$. We can call this percentage decrease that A yields in the production of E per se, A's *overall preventive influence* $OPI_{A,E}$ on E. The overall preventive influence is a two-place relation between A and E and is not relative to any specific generative cause of E. Its value is highly contextual, since it not only depends on the intrinsic cause-specific preventive powers of A relative to each generative cause of E, but also on the probabilities of each generative cause of E, as well as on their generative causal powers.

5.6.3 Relative Risk Reduction

A discipline in which measures of preventive power play an important role is epidemiology. A common measure that is used to determine the efficacy of a treatment or a drug after a clinical trial is called *relative risk reduction*. Imagine a clinical trial, in which we aim to determine the effectiveness of a drug A against a certain symptom E. In an experimental group (1), all individuals are administered with A. In a control group (2), it is made sure that no individual takes A.

Group	E	$\neg E$
(1)	a	b
(2)	c	d

Table 5.1: Clinical Trial Results. Adapted from (Stegenga, 2015, p. 66).

With the results represented in table 5.1, we can determine the proportion of individuals with E in each group:

- Experimental Event Rate (EER) $= \dfrac{a}{a+b}$
- Control Event Rate (CER) $= \dfrac{c}{c+d}$

The relative risk reduction (RRR) is defined in the following way:[10]

$$RRR = \frac{CER - EER}{CER} \tag{5.20}$$

Since $P(E|A) = EER$ and $P(E|\neg A) = CER$, the relative risk reduction RRR is nothing else than the overall preventive influence of A on E as measured by $f_{ch-prev}(A,E)$. As a consequence, RRR can, just like $f_{ch-prev}(A,E)$, in general not determine the cause-specific intrinsic preventive power(s) of A on E. But very often, it is the cause-specific preventive power that is especially valuable for making medical decisions. Imagine that E is some unpleasant symptom, C is some

[10] See, for example, (Stegenga, 2015, p. 66).

virus that produces E with the causal power $g_{C,E} = 0.5$ and A is a medicine that can prevent C from producing E with the preventive power $h_{A(C),E} = 0.6$. Now imagine a clinical trial, in which the effectiveness of A on preventing E is supposed to be tested. The test group and the control group both consist of individuals that all have C. We therefore have $P(C) = 1$ in both groups. In the test group each individual is administered with A, while in the control group no one receives A. But there is an independent background cause B of E that cannot be eliminated. B is some genetic disposition that can produce E with the causal power $g_{B,E} = 0.8$ and B is evenly distributed with $P(B) = 0.5$ in both groups. The example has the exact same structure and the same probabilities as the causal scenario, that we have considered in the previous section and for which we determined the value of $f_{ch-prev}(A, E)$ in equations (5.17)-(5.19). We therefore have: $RRR = f_{ch-prev}(A, E) = 0.256$. Now imagine Peter, who has some interest in avoiding symptom E and who has just been infected with virus C. Peter knows that he does not have the genetic predisposition B. He considers to take medicine A, but is unsure, because it has several side-effects. So, he would only take it, if it significantly lowers the probability of getting symptom E. Clearly, the value of RRR is misleading for Peter. The information he needs is the C-specific preventive power of A on E, which is 0.6, and not the overall preventive influence of A on E in a population that not only has the virus C but also, at least partially, the genetic predisposition B.

Notice that for the empirical determination of $h_{A(C),E}$ two clinical trials are needed. First, we need a clinical trial, in which the intrinsic causal power $g_{C,E}$ of C on E is determined. With $P(A) = 0$, $P(B) = 0.5$, $P(C) = 1$ in the test group and $P(A) = 0$, $P(B) = 0.5$, $P(C) = 0$ in the control group, $f_{ch}(C, E)$ can be used to determine $g_{C,E}$. We then need a clinical trial, in which C's effective causal influence on E in the presence of A, that is $ECI^{A}_{C,E}$, is determined. With $P(A) = 1$, $P(B) = 0.5$, $P(C) = 1$ in the test group and $P(A) = 1$, $P(B) = 0.5$, $P(C) = 0$ in the control group, $f_{ch}(C, E)$ can again be used to determine $ECI^{A}_{C,E}$. With these two values, we can calculate the percentage decrease from $g_{C,E}$ to $ECI^{A}_{C,E}$, which gives us the value of $h_{A(C),E}$.

5.7 Absolute vs. Relative Outcome Measures

In chapter 4, I have already pointed out that all different measures of causal strength can be broadly divided into two categories: absolute and relative outcome measures. Absolute outcome measures, like $CS_e(C, E)$ or $CS_s(C, E)$, measure the absolute difference that introducing C would make for the probability of E in comparison to some alternative action, like introducing $\neg C$ or doing nothing at all. Relative outcome measures, like $CS_{ch}(C, E)$, on the other hand, measure the difference that introducing C would make for the probability of E in terms of a percentage increase or decrease of E's probability. The same differentiation between absolute and relative outcome measures can be carried over to measures of preventive strength. Our measure CS_{prev} of intrinsic, cause-specific preventive power and the measure of overall preventive influence, as determined by $f_{ch-prev}$/RRR, are both relative outcome measures. But we could also use an analogue of Eells' measure CS_e, namely $CS_{e-prev}(A, E) = P(E|\neg A) - P(E|A)$, to measure the preventive strength of an event A on an effect E. $CS_{e-prev}(A, E)$ is an example of an absolute outcome measure when it comes to measuring preventive strength.[11]

Stegenga (2015), as well as Sprenger and Stegenga (2017), argue that we should generally prefer absolute outcome measures over relative outcome measures when it comes to making decisions based on causal information. Unlike the arguments by Sprenger (2018), which I discussed in chapter 4, these arguments do not entail the claim that relative outcome measures are completely inadequate as measures of causal strength. Instead, they only suggest that relative outcome measures are significantly less useful for decision making than absolute outcome measures and should therefore be avoided for this purpose. Even though this does not include a complete rejection of relative outcome measures, it advocates for a serious restriction of their applicability.

Stegenga and Sprenger aim to establish four different claims with their arguments. (1) Knowing the value of a relative outcome measure is not enough to enable a rational choice between alternative actions. (2) The value of a relative outcome measure is potentially misleading and can therefore lead to decisions that are not optimal according to ones own beliefs and values. (3) Knowing the value of an absolute outcome measure is sufficient for making a rational choice between alternative actions. (4) Absolute outcome measures are not misleading and do therefore not lead to decisions that are suboptimal according to ones own beliefs and values.

I agree with (1) and (2). But I will argue that this does not establish the claim that relative outcome measures are inferior to absolute outcome measures when it comes to decision making. One reason for this is that claims (3) and (4) are wrong. I will argue that absolute outcome measures are typically not sufficient for making rational decisions and that they are potentially misleading as well. I will additionally argue that, while relative outcome measures are indeed not sufficient for rational decision making, certain relative outcome measures are necessary for it. The same does not hold for any absolute outcome measure.

5.7.1 Relative Outcome Measures Are Insufficient For Rational Decisions

Let us start with Stegenga and Sprenger's argument for (1) and (3). Imagine you are offered a certain drug A that has proved to be effective in preventing a certain unpleasant symptom E.[12] A common method for making a rational choice between taking A or not taking A is maximizing the expected utility. As Sprenger and Stegenga (2017, p. 846) point out, the expected utility of taking the drug A is given by the following formula:

$$EU(A) = P(E|A) \times u + P(\neg E|A) \times u' - a \tag{5.21}$$

where a is the cost of taking A, u is the utility/cost of E, and u' is the utility/cost of $\neg E$. Here again, $P(E|A)$ and $P(\neg E|A)$ should not be understood as conditional probabilities, but as the probabilities of E and $\neg E$ after realizing A by an intervention. Likewise, we have:

[11] Similar to the measure of overall preventive influence, $CS_{e-prev}(A, E)$ treats prevention as a two-place relation between a preventer A and an effect E.

[12] As argued above, we should actually say: A prevents some or several generative causes of E from producing E. But Sprenger and Stegenga (2017) do not treat prevention as a three-place relation. In their arguments, Sprenger and Stegenga (2017) therefore only use RRR as representing relative outcome measures. I have pointed out in the previous section that RRR is not always equatable with CS_{prev}. Nonetheless, Sprenger's and Stegenga's arguments can easily be adapted for CS_{prev} as well.

$$EU(\neg A) = P(E|\neg A) \times u + P(\neg E|\neg A) \times u' - b \tag{5.22}$$

where b is the utility/cost of not taking A. After a few algebraic transformations, we get:

$$EU(A) > EU(\neg A) \text{ if and only if } |P(E|A) - P(E|\neg A)| > \frac{a-b}{u-u'} \tag{5.23}$$

As Sprenger and Stegenga (2017, p. 846) point out, this shows that as soon as the utilities a, b, u, u' are known, the absolute outcome measure $|P(E|A) - P(E|\neg A)|$ is sufficient for making a rational decision between A and $\neg A$.[13] The same does not hold for a relative outcome measure. Sprenger and Stegenga (2017, p. 847) show that the following holds:

$$EU(A) > EU(\neg A) \text{ if and only if } |P(E|\neg A) > \frac{a-b}{(u-u') \times RRR} \tag{5.24}$$

This shows that, if we know the utilities a, b, u, u', knowing the value of the relative outcome measure RRR is not sufficient for a rational choice between A and $\neg A$. We additionally need to know the value of $P(E|\neg A)$.

5.7.2 Relative Outcome Measures Can Be Misleading

Relative outcome measures are not only insufficient for making rational decisions, they can also be misleading. Stegenga points out: "Relative measures, by promoting the base rate fallacy, fundamentally mislead patients into overestimating effectiveness" (Stegenga, 2015, p. 68). In a similar vein, Broadbent (2013) remarks: "Some commentators have suggested that risk relativism is a tool of exaggeration, making the small seem huge and the unimportant important, perhaps for financial gain" (Broadbent, 2013, p. 130).[14] And Sprenger and Stegenga (2017) point out that "[i]t is a robust empirical finding that physicians are more likely to prescribe a drug when the risk is expressed in relative than in absolute terms" (Sprenger and Stegenga, 2017, p. 844). To illustrate how this could happen, consider the following imaginary example.[15] A drug A is supposed to lower the risk of a symptom E. In a clinical trial that aims to test the efficacy of A, every individual in the test group is administered with A, while every individual in the control group is administered with a placebo. The results are the following: 2% of the individuals in the control group developed E, while only 1% of the individuals in the test group developed E. When we use the absolute outcome measure $CS_{e-prev} = P(E|\neg A) - P(E|A)$, the efficacy of A is indicated as 0.01: Taking medicine A changes the value of $P(E)$ by 1%. But if we use the relative outcome measure RRR, then the efficacy of A is indicated as 0.5: Taking medicine A decreases the value of $P(E)$ by 50%.

To be clear: Both values are correct. They simply describe two different things. It is correct that, after taking A, the risk of developing E has decreased by 50% compared to its previous value. But the risk of developing E was quite small in the first place, namely 2%. This is why A prevented E in only 1% of all the individuals in the test group. In another 1% of the individuals,

[13] Notice that $|P(E|A) - P(E|\neg A)| = CS_e(A, E)$ if A is a generative cause of E and $|P(E|A) - P(E|\neg A)| = CS_{e-prev}(A, E)$ if A is a preventer of E.

[14] Broadbent uses the term 'risk relativism' to denote a preference of relative over absolute outcome measures.

[15] See (Stegenga, 2015) for some real-life examples.

A failed to prevent E, because E developed anyhow. And in 98% of all the individuals in the test group, A was completely useless, because nothing causally produced E in the first place. So, only 1% of the individuals actually benefited from A. The problem with the value of the relative outcome measure RRR is not that it is false. People, doctors and patients alike, simply tend to confuse the meaning of a relative outcome measure with the meaning of an absolute outcome measure. Learning that A has an RRR-value of 50% seems often to be confounded with the idea that in 50% of the patients A, if administered, would prevent E. As Sprenger and Stegenga (2017) point out, this confusion even seems to persist, when the base rate of E, in our example 2%, is explicitly stated.

These facts prompt Stegenga (2015), as well as Sprenger and Stegenga (2017), to declare a general superiority of absolute outcome measures over relative outcome measures when it comes to decision making.

5.7.3 Relative Outcome Measures Are Necessary For Decisions

I partially agree with Sprenger and Stegenga's arguments. It is true that knowing the value of a relative outcome measure is not sufficient for making a rational decision between two alternative actions. And the fact that describing the efficacy of a drug in terms of a relative outcome measure often leads to an overestimation of its usefulness is clearly problematic. But, I do not agree that all this establishes the superiority of absolute outcome measures over relative outcome measures for the purpose of decision making, especially not over the relative outcome measure CS_{prev}. Just take the example from the previous section. The absolute outcome measure $CS_{e-prev}(A, E) = P(E|\neg A) - P(E|A) = 0.02$ does indeed give us some very useful information about the context of the clinical trial, namely the probability that A, if administered, would prevent E in the test group of the clinical trial. But this probability is typically not extrapolatable to other populations or individuals. As Stegenga himself points out:

> "Subjects in a clinical trial are virtually never drawn from a random sample of the broader population who have the disease in question, and the criteria that determine eligibility for a clinical trial often render subjects in a trial different in important respects from the broader population of people who have the disease" (Stegenga, 2015, p. 69).

Now, imagine Suzy, who has to decide whether she should take A to prevent E. Suzy works in a hospital and she knows that her risk of developing E is much higher than the risk for the popoulation in the clinical trial. E might for example be caused by certain viruses that are transmitted through contact with infected people. The values of $P(E|A)$ and $P(E|\neg A)$, which hold in the clinical trial, are therefore not extrapolatable to Suzy's situation. As a consequence, the value of $CS_{e-prev}(A, E) = 0.02$ is not extrapolatable to Suzy's situation. Suzy can therefore not use the value of $CS_{e-prev}(A, E)$ to determine whether taking A or whether not taking A is more rational for her, that is, whether $EU(A) > EU(\neg A)$. The value of $CS_{e-prev}(A, E)$, which describes the probability that A, if administered, prevents E in the trial-population, does not describe the probability that A, if Suzy takes it, prevents Suzy from developing E. The only chance of determining this probability is by learning the intrinsic causal powers of the generative causes of E, the probabilities that the generative causes of E are present in Suzy's situation, and the intrinsic cause-specific preventive powers of A on E. As previously argued, the relative outcome measures CS_{ch} and CS_{prev} are the best candidates for measuring intrinsic causal powers. If, for example, there is one generative cause C of E, then the probability that A, if Suzy takes it, prevents Suzy from developing E, is given by the following formula:

$$\begin{aligned} P(E|\neg A) - P(E|A) &= P(C) \times CS_{ch}(C,E) - [P(C) \times CS_{ch}(C,E) \times (1 - CS_{prev}(C(A),E))] \\ &= P(C) \times CS_{ch}(C,E) \times CS_{prev}(C(A),E) \end{aligned} \tag{5.25}$$

It is the result of this formula that should form the basis for Suzy's decision. The formula contains more than just the value of a relative outcome measure, since it also contains the base rate $P(C)$ of C. But unlike the value of the absolute outcome measure $CS_{e-prev}(A,E)$ as it has been determined in the clinical trial, the values of the relative outcome measures $CS_{ch}(C,E)$ and $CS_{prev}(C(A),E)$ are extrapolatable to new contexts and therefore serviceable for Suzy.

I am not saying that the respective intrinsic causal powers are easy to determine. It actually needs several clinical trials to do so. But only intrinsic causal powers are extrapolatable from learning contexts, like clinical trials, to application contexts, like Suzy's decision. Values of causal strength that are not extrapolatable from clinical trials to real-life situations are of no use for any decision in a real-life situation.[16] Absolute outcome measures are therefore usually not sufficient for making rational decisions between alternative actions, for the simple reason that the values of absolute outcome measures are typically not extrapolatable to the contexts, in which the decisions are due.[17] Since only the values of the relative outcome measures CS_{ch} and CS_{prev} are extrapolatable to new situations, these measures are actually indispensable for real-life decision making.

There is another respect, in which absolute outcome measures can be insufficient and misleading for rational decisions. The method of maximizing the expected utility as used by Sprenger and Stegenga (2017) does not consider the possibility of risk aversion. But especially in medical decisions, risk aversion is an important factor. It makes a significant difference, whether taking A reduces the risk of dying from 1% to 0% or from 81% to 80%. In both cases, the absolute outcome measure CS_{e-prev} ascribes A a preventive strength of 0.01 on death. But in the first scenario, taking A seems to make a much more valuable impact, because it completely removes a non-negligable risk of dying. The reduction from 81% to 80%, on the other hand, does, intuitively, not make a very significant difference. So, even if we could extrapolate the value of

[16] Stegenga (2015) also proposes some potential solutions to the extrapolation problem. One proposal is to aim for knowledge about the mechanism of the causal prevention at work. This proposal is closely related to the point that I have just made. What is crucial about the mechanism of prevention is which generative causes the preventive cause can actually prevent and which of those generative causes are actually present in a given situation.

[17] Notice that the same accusation holds for all relative outcome measures besides CS_{ch} and CS_{prev}. As shown above, RRR is also highly contextual and therefore typically not extrapolatable. Broadbent (2013, p. 139 - 40) actually points out that, if a relative outcome measure would be better extrapolatable than absolute outcome measures, then its preference would be justified. He then points out that relative outcome measures like RRR are just as contextual as absolute outcome measures. Broadbent is right about RRR. But he does not consider CS_{ch} and CS_{prev}.

$CS_{e-prev}(A, E)$ to the situation of interest, it would not be sufficient for making a rational decision, as soon as risk aversion place a role. Just like with relative outcome measures, we would additionally need to know the base rate of the event E, that we aim to prevent.

5.7.4 Summary

Stegenga and Sprenger are right when they claim that (1) knowing the value of a relative outcome measure is not enough to enable a rational choice between alternative actions, and that (2) the value of a relative outcome measure is potentially misleading and can therefore lead to decisions that are not optimal according to ones own beliefs and values. But both points equally hold for absolute outcome measures. Absolute outcome measures are typically not extrapolatable from the clinical trials, in which they are determined, to real-life contexts, in which decisions are due. They additionally yield misleading evaluations of the impact of potential actions, whenever risk aversion place a role. But unlike any absolute outcome measure, the relative outcome measures CS_{ch} and CS_{prev}, are typically indespensible for making rational decisions in real-life situations, since only the values of intrinsic causal powers can be extrapolated from learning contexts, like clinical trials, to the application contexts of real-life situations.

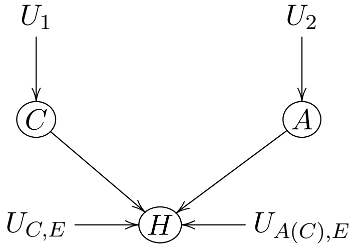
So, while relative outcome measures do indeed have weaknesses for the purpose of decision making, these weaknesses cannot be overcome by simply rejecting the use of relative outcome measures altogether and instead relying on absolute outcome measures only. We cannot spare the use of relative outcome measures, like CS_{ch} and CS_{prev}, for the purpose of decision making. They only have to be used correctly, which includes the avoidance of the base rate fallacy.

5.8 Preventive Power in Causal Models

Having explicated a measure CS_{prev} of intrinsic, cause-specific preventive power, we now face the same question as we did after the explication of intrinsic generative causal power: Where and how do values of intrinsic, cause-specific preventive power feature in probabilistic SEMs?

For generative causes with a non-maximal causal power on their effects, we have already introduced bivalent error-terms into probabilistic SEMs. We have pointed out that for a cause candidate $C = 1$ of $E = 1$, the corresponding error-term $U_{C,E}$ taking on the value 1 is supposed to represent, whether $C = 1$ would successfully produce $E = 1$, if $C = 1$ would be the case. As pointed out in section 5.5, this interpretation needs a slight adjustment, as soon as we consider the possibility of preventive causes. $U_{C,E} = 1$ has to be interpreted as saying that $C = 1$ would successfully produce $E = 1$, if $C = 1$ would be the case and no C-specific preventive cause of E would be present.

For preventive causes with non-maximal preventive powers, we can proceed similarly as with generative causes and introduce error-terms into SEMs to represent the probabilistic causal relationships of prevention. For a $(C = 1)$-specific preventive cause $A = 1$ of $E = 1$ with a $(C = 1)$-specific preventive power $h_{C(A),E}$ on E, we can introduce the bivalent error-term $U_{A(C),E}$. $U_{C(A),E} = 1$ represents the fact that $A = 1$ would prevent $C = 1$ from producing $E = 1$, if $A = 1$, $C = 1$ and $U_{C,E} = 1$ would be the case. Consider a simple example with three bivalent variables C, E and A, with $C = 1$ being a generative cause of $E = 1$ with the generative causal power $g_{C,E}$ and $A = 1$ being a $(C = 1)$-specific preventive cause of $E = 1$ with the power $h_{C(A),E}$. The scenario can be represented by the causal model presented in figure 5.1.



- $C := U_1$
- $A := U_2$
- $E := (C \wedge U_{C,E}) \wedge \neg(A \wedge U_{A(C),E})$

Figure 5.1: Simple prevention.

Imagine that the SEM is supposed to represent a past token scenario. With our interpretation of the error-term $U_{A(C),E}$ it is clear that, as long as we have no trumping evidence about the events causally downstream of C and A, $h_{C(A),E}$ amounts to the best guess for the value of $\mathcal{P}(U_{A(C),E} = 1)$, just as $g_{C,E}$ is our best guess for the value of $\mathcal{P}(U_{C,E} = 1)$.

In the previous formulations of PROBAC and PROSAC, we have not yet explicitly considered the presence of preventive causes. We can now correct this neglect by once again amending the first step of both guidelines:

(1) Create a preliminary probability distribution $\mathcal{P}_{pre}$ such that:

  (a) $\mathcal{P}_{pre}$ ascribes probabilities to the background variables in $\mathcal{M}$ that are consistent with the current knowledge about the token scenario.

  (b) If $U_{X,Y}$ is an error-term that represents, whether $X = x$ causally produces $Y = y$, given $X = x$ and the absence of any $(X = x)$-specific preventive cause of $Y = y$, then $\mathcal{P}_{pre}(U_{X,Y} = 1) = CS_{ch}(X = x, Y = y)$.

  (c) If $U_{Z(X),Y}$ is an error-term that represents, whether $Z = z$ prevents $X = x$ from producing $Y = y$, given $X = x$, $Z = z$ and $U_{X,Y} = 1$, then $\mathcal{P}_{pre}(U_{Z(X),Y} = 1) = CS_{prev}(Z = z(X = x), Y = y)$.

But here again, not in every probabilistic SEM is the value of $\mathcal{P}(U_{Z(X),Y} = 1)$ equal to the $(X = x)$-specific preventive power of $Z = z$ on $Y = y$. As soon as we have updated the probability distribution $\mathcal{P}$ in the probabilistic SEM by conditionalizing on some trumping information about events that lie causally downstream from $X = x$ or $Z = z$, the value of $\mathcal{P}(U_{Z(X),Y} = 1)$ may very well differ from $CS_{prev}(Z = z(X = x), Y = y)$. Consider the simple prevention scenario from figure 5.1 and imagine that it represents a token scenario, about which we know that $\mathcal{P}(C = 1) = 1$, $\mathcal{P}(A = 1) = 1$, $g_{C,E} = 1$, and $h_{A(C),E} = 0.5$. As long as we have no further evidence about the actual value of E, we have: $\mathcal{P}(U_{C,E} = 1) = 1$, $\mathcal{P}(U_{A(C),E} = 1) = 0.5$, and $\mathcal{P}(E = 1) = 0.5$. But now imagine, we somehow come to learn that $E = 0$. As soon as we update $\mathcal{P}$ by conditionalizing on $E = 0$, we get a new distribution $\mathcal{P}'$ with: $\mathcal{P}'(U_{A(C),E} = 1) = 1$. By learning that $E = 1$ did not happen, we also learn that $A = 1$ must have prevented $C = 1$ from producing $E = 1$. This knowledge is expressed in the new value of $\mathcal{P}'(U_{A(C),E} = 1)$, which is 1. But this value is not the intrinsic ($C = 1$)-specific preventive power of $A = 1$ on $E = 1$. The preventive power remains to be 0.5, even though we now know that $A = 1$ did actually prevent $C = 1$ from producing $E = 1$ in the given situation.

5.9 Summary

In the present chapter, I have argued that, if we accept the intuition of difference, we cannot simply reduce prevention to production of the complement and, consequently, preventive power to generative causal power on the complement. To respect the intuition of difference, I have instead expanded Cheng's Power PC theory by some intuitive assumptions about the concept of prevention. This expansion of the Power PC theory allowed us to deduce a measure CS_{prev} of intrinsic preventive power that treats prevention as a three-place relation. I have shown that this measure is, in general, distinct from well known measures of preventive power that have been put forward in the literature, especially from Cheng's measure $f_{ch-prev}$ and from the measure of relative risk reduction (RRR), which is well entrenched in epidemiology. Both these measures do not measure the intrinsic preventive power of a preventive cause, but rather the contextual overall preventive influence.

I have further defended the two measures of intrinsic causal power, CS_{ch} and CS_{prev}, against Sprenger and Steganga's arguments, according to which relative outcome measures are in general inferior to absolute outcome measures for the purpose of decision making. I have argued that CS_{ch} and CS_{prev} are typically indispensable for making decisions in real-life scenarios, while absolute outcome measures have just the same weaknesses as relative outcome measures for the purpose of decision making: Their values are typically insufficient and potentially misleading for rational decisions.

Chapter 6

Production and Prevention

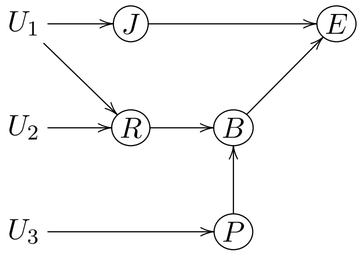
6.1 Introduction

In the last two chapters I have expanded Cheng's Power PC theory to obtain two measures of intrinsic causal power, generative and preventive, that are more generally applicable than Cheng's original measures. We have already seen how quantities of intrinsic causal power can be used in the construction of probabilistic SEMs. This illustrates that our expansion of the Power PC theory provides us with concepts that are not only reconcilable with the causal model framework, but that even form valuable components of it. But in my investigation on how the Power PC theory and the causal model framework fit together, a crucial question has remained unanswered so far. How do the concepts of causal production and prevention, both concepts of token causation, that lay at the heart of the Power PC theory, correspond to the concepts of token causation, like actual and strong actual causation, that we have defined in the causal model framework?

In the present chapter, I aim to answer this question. To do so, I will have a closer look at the theoretical concepts of production and prevention that were taken as primitives in the previous two chapters and that were only characterized by means of some intuitive postulates. The comparison of these concepts with the interventionist concepts of actual, sufficient, and strong actual causation will not only lead us to a more comprehensive understanding of causation. It will ultimately also enable a more thorough understanding of causal explanation.

6.2 Production and Dependence: Two Conceptions of Causation

To gain an impression of how the concept of causal production relates to the interventionist concepts of token causation, as defined in chapters 1 and 2, consider the following scenario: Michael Jordan is about to shoot from the three point line. Magic Johnson saw what Jordan was up to and started to run towards Jordon to get in front of him and block his throw. But Johnson does not reach Jordan. Instead he bumps into Scottie Pippen, who just happened to cross Johnson's path to Jordan. So, Johnson does not block Jordon's throw and the ball reaches the basket undisturbed and goes through the hoop. We can represent the scenario by the structural equation model $\mathcal{M}^J$ that is shown in figure 6.1 and in which J (representing whether Jordan throws), R (representing whether Johnson starts to run towards Jordon), B (representing whether Johnson blocks Jordan's throw), P (representing whether Pippen crosses Johnson's path towards Jordan), and E (representing whether Jordan scores) are bivalent variables with the value 1 representing that the corresponding event takes place and the value 0 representing that the corresponding event does not takes place.[1]



- $J := U_1$
- $R := U_2 \wedge U_1$
- $P := U_3$
- $B := R \wedge \neg P$
- $E := J \wedge \neg B$

Figure 6.1: $\mathcal{M}^J$ - the basketball scenario.

The given causal setting for $\mathcal{M}^J$ is $U_1 = U_2 = U_3 = 1$, which gives us: $J = 1$, $R = 1$, $P = 1$, $B = 0$, $E = 1$.

Intuitively, Jordan's throw ($J = 1$) causally produces the score ($E = 1$), while Pippen's crossing of Johnson's path to Jordan ($P = 1$) prevents Johnson from blocking Jordan's throw ($B = 1$). If the block ($B = 1$) would have happened, it would have prevented Jordan's throw ($J = 1$) from producing the scoring ($E = 1$). This is why Pippen's crossing of Johnson's path to Jordan ($P = 1$) serves as a double-preventer: it prevented the causal production of an event ($B = 1$) that would have prevented the causal production of the scoring ($E = 1$). But intuitively, neither Pippen's crossing of Johnson's path to Jordan ($P = 1$) nor the omission of Johnson's block ($B = 0$) causally produces the score ($E = 1$)

Applying the HP-definition of actual causation, we obtain that $J = 1$ is an actual cause of $E = 1$, $B = 0$ is an actual cause of $E = 1$, and $P = 1$ is an actual cause of $E = 1$.[2] This shows that the concept of actual causation is much more permissive than the concept of causal production. Not only the causal producer, namely Jordan's throw, is an actual cause of the scoring. The omission of a preventer, namely the omission of Johnson's block, is also an actual cause of the scoring, as is the double preventer, namely Pippen's crossing of Johnson's path to Jordan. Applying Halpern's definition of sufficient causation, we obtain that the conjunction of Jordan throwing and Johnson not blocking ($J = 1 \wedge B = 0$), as well as the conjunction of Jordan throwing and Pippen crossing Johnson's path ($J = 1 \wedge P = 1$) are sufficient causes of the scoring. So, in the given situation, any sufficient cause of the scoring contains not only the causal producer of the scoring, but also an omission of a preventer of the scoring or a preventer

[1] $\mathcal{M}^J$ is clearly a simplification, since we pretend that all causal relationships in the scenario are deterministic. For the following arguments, this simplification is harmless, since they hold for probabilistic causal relationships just as well as for deterministic causal relationships.

[2] In all three cases, we have a simple counterfactual dependence of $E = 1$ on the respective cause.

of a preventer of the scoring.

The example nicely illustrates that, even though the concept of causal production may sometimes coincide with interventionist concepts of token causation, like actual or sufficient causation, it is intensionally and extensionally clearly distinct from these concepts. This result conforms with Ned Hall's (2004) famous thesis that there are two decisively distinct conceptions or interpretations of causation, which he labels *dependence* and *production*. The dependence-interpretation of causation is based on the idea that all there is to causation is some kind of counterfactual dependence between cause and effect. This does typically not amount to the claim that a simple counterfactual dependence of the effect on the cause is both necessary and sufficient for causation. Instead, Dependence-accounts of causation typically take, what Yablo (2002) calls, 'de facto dependence' as a condition for causation, which means that the effect counterfactually depends on the cause in a certain contingency or under the assumption that certain factors are kept fixed by interventions. The definitions of actual, sufficient, and strong actual causation, as presented in chapters 1 and 2, all comply with the dependence-interpretation of causation.[3] According to the production-conception of Causation, on the other hand, there is something more to causation than a mere de facto counterfactual dependence of the effect on the cause. We still have to find out what exactly this additional something is. But in the philosophical literature, it has come to be known under the somewhat sketchy and figurative name 'biff'.[4]

The differentiation between a dependence- and a production-interpretation of causation accords with the widely shared intuition, that some cases of causation appear to be more genuine or prototypical than other cases of causation. It seems fair to say that Jordan's throw, the omission of Johnson's block, and Pippen's crossing of Johnson's path to Jordan are all causes of the scoring. The scoring is (de facto) dependent on all three events. But among the three causes, Jordan's throw clearly stands out. It is Jordan's throw that actively produces the balls momentum and thereby puts it on its scoring path. It is Jordan's throw that has *biff*. This explains why the scoring will only appear on Jordan's list in the game-statistics. It will not be attributed to either Johnson or Pippen. Hall summarizes the intuition like this: "production does seem, in some sense, to be the more 'central' causal notion" (Hall, 2004, p. 256).[5]

Although our intuition seems to put a spotlight on the significance of a causal producer and thereby sets it apart from mere dependence-causes of the given effect, intuition becomes surprisingly blurred when it comes to explicating what a causal producer really is and in what exactly it differs from mere dependence-causes. The two accounts of causation, that we have presented in this dissertation so far, are of no real help either. The interventionist account of causation, as presented in chapters 1 and 2, provides us with necessary and sufficient conditions

[3] Hall (2004) actually speaks of "two concepts of causation", when he differentiates *dependence* from *production*. But I consider this formulation to be slightly misleading. At least when it comes to *dependence*, we are not dealing with a single concept of causation, but with a whole group or family of concepts, like actual, sufficient, and strong actual causation. This is why I rather speak of two distinct interpretations or conceptions of causation.

[4] See, for example, (Lewis, 2004), (Illari, 2013), or (Handfield et al., 2008).

[5] Hall is not the only one who differentiates between a dependence- and a production-conception of causation. Nor is he the only one, who considers the production-conception to be the more central or prototypical causal notion. For example, Glennan (2017), Strevens (2008), and Dowe (2000) hold similar positions. As we will see in chapter 9, Woodward (2015) puts forward a very similar differentiation between "thin" and "thick" concepts of causation.

for identifying certain dependence-causes, like actual, sufficient or strong actual causes. But it does not yet tell us how to identify a causal producer. The amended Power PC theory, as developed in chapters 4 and 5, does not fare much better. Although the theory specifies certain features of causal production that help us to determine a method for measuring the power of a given producer on its effect, these features are not sufficient to determine whether something is a producer in the first place. This is why we have treated the concept of causal production, just like the concept of prevention, as a basic, unanalyzed, theoretical term. Theories that aim to explicate, what causal production really is, are rare and often problem-ridden. Illari (2013) gives the following assessment of the situation:

> "The majority of work is still on difference-making [dependence] accounts, with the literature on each type of difference-making account being vast. Correspondingly, the literature on production is a drop in a counterfactual ocean. Further, existing production accounts have recognized weaknesses, and few seem inclined to try to fix them" (Illari, 2013, p. 97).

Hall, who considers production "to be the more 'central' causal notion", also admits that production is "rather more difficult to characterize" (Hall, 2004, p. 225). This is why Hall does not ground his differentiation of production from dependence on a full-flegded account of causal production. Instead he only puts forward three features that he considers to hold generally for causal production, but not for a dependence-conception of causation:[6]

- *Transitivity*: If event B causally produces event C and C causally produces event E, then B causally produces E.
- *Intrinsicness*: Whenever an event C causally produces event E, then this is due to something intrinsic to C and E.
- *Locality*: Whenever C causally produces E, there is some spatiotemporal connection or link between C and E.

Two of the three features clearly match with our characterization of causal production in our generalized Power PC theory. We have explicitly assumed that transitivity holds for causal production in axiom C6 and since the Power PC theory ascribes intrinsic causal powers to causal producers, intrinsicness is covered as well.[7] But so far, we have not made any kind of locality-assumption about causal production in our expanded Power PC theory.

Hall is not alone in considering locality to be a crucial feature of causal production. In fact, locality plays the dominant role in the most well-known explications of causal production in the literature. It seems to be the main bearer of hope, when it comes to finding a criterion, which enables us to differentiate causal producers from mere dependence-causes. But in the following sections, I will argue that locality is at least disputable as a general feature of causal production and that we will likely be disappointed, if we put all our hope on locality as a distinguishing feature of causal production.

[6] See (Hall, 2004, p. 225).

[7] Hall explicates Intrinsicness somewhat differently though. See (Hall, 2004, p. 239) for his explication, which I consider to be problematic when it comes to probabilistic causal relationships. But a discussion of this would go beyond the scope of this chapter and is irrelevant for the following discussion.

6.3 Existing Accounts of Production-Causation

6.3.1 Process Accounts

Natural candidates for explications of our concept of causal production are the so-called process accounts of causation. Process accounts of causation do not understand causation in terms of a de facto counterfactual dependence, but as a result of certain objects transferring conserved quantities when their world-lines interesect. This idea amounts to a physical explication of Hall's locality thesis. The two most prominent process accounts are the mark transmission (MT) theory by Salmon (1984) and the conserved quantity (CQ) theory by Dowe (2001). Fair (1979) can be seen as a precursor to both theories. Later on, Salmon (1997) abandoned his original MT theory, mainly due to the fact that his definitions of *causal processes* and *causal interactions* crucially depend on counterfactuals. Salmon's revisions of his theory converge to Dowe's CQ theory.[8] Although there are still some slight differences, I will confine myself in in the following discussion to Dowe's CQ theory which, according to Hitchcock, is "[b]y far the most developed attempt" (Hitchcock, 2009b) among the process theories to analyze the concept of causation.

The core of the CQ theory is given by the following two definitions:[9]

CQ1 A *causal process* is a world line of an object that possesses a conserved quantity.

CQ2 A *causal interaction* is an intersection of world lines that involves exchange of a conserved quantity.

Dowe further explicates: "A world line is the collection of points on a spacetime (Minkowski) diagram that represents the history of an object" (Dowe, 2000, p. 90), and: "An object is anything found in the ontology of science (such as particles, waves and fields), or common sense (such as chairs, buildings and people). [...]. A conserved quantity is any quantity that is governed by a conservation law, and current scientific theory is our best guide as to what these are" (Dowe, 2000, p. 91). Typical examples are energy, linear momentum, angular momentum, or charge. Not every process is causal, though. According to the CQ theory, any world line of an object is a process. So, whenever the object in question does not possess any conserved quantity, the process is non-causal, a so called *pseudo process*, that cannot contribute to the causal production of anything. An example of such a pseudo process is the movement of a shadow. According to Dowe, a shadow classifies as an object that can have properties, like having a certain shape or moving at a certain velocity. But the shadow does not possess a conserved quantity, like energy or linear momentum.

To clarify the definition of a causal interaction, we still need some further elaboration on what is meant by an intersection and by an exchange of conserved quantities. Dowe explains: "An *intersection* is simply the overlapping in spacetime of two or more processes. [...] An *exchange* occurs when at least one incoming, and at least one outgoing process undergoes a change in the value of the conserved quantity" (Dowe, 2000, p. 91-92. *Emphasis in original*). A simple example is the collision of two billiard balls. Before the collision, both balls possess a

[8] For a comparison of Salmon's revised theory or Fair's transference theory with Dowe's CQ theory, see (Dowe, 2001, p. 109 ff.).

[9] See (Dowe, 2000, p. 90).

certain value of linear momentum. After the collision, that is the intersection of the world lines of the balls, each ball possesses a different linear momentum than before. So, by CQ2, a causal interaction took place.

The CQ theory seems to enable us to differentiate between causal and non-causal processes or interactions.[10] But so far, the theory does not say anything about the relation of causal production, whose relata are events. How can we employ the concepts of causal processes and causal interactions to analyze the relation of causal production? Here comes Dowe's tentative answer, which he labels the naive process theory:[11]

Naive Process Theory (NPT). *C causally produces E if and only if a continuous line of causal processes and causal interactions obtains between C and E.*

Dowe does not explicate how exactly events can be connected by a continuous line of causal processes and causal interactions. But if we assume that events consist of certain objects which possess or change certain properties at a certain time or time interval, then it follows that events can contain causal processes and causal interactions. A reasonable interpretation of NPT is therefore that C causally produces E if and only if C contains a causal process or a causal interaction that is the first element of a continuous line of causal processes and causal interactions, which eventually leads to a causal process or a causal interaction that is contained in E. So, if C contains the causal process p_1, while E contains the causal process p_n, then there must be causal processes p_2, p_3, ..., p_{n-1}, such that p_i interacts with p_{i+1} for all $1 \leq i \leq n-1$.

NPT is clearly an explication of the locality thesis. So, how does it fare as an analysis of causal production? At a first glance, it seems to do well with prototypical examples of causal production, like the collision of billard balls or the smashing of bottles. Especially in situations of preemption or overdetermination, which proved to be quite hard to handle for counterfactual theories of causation, NPT seems to shine. When Suzy and Billy both throw a stone at a bottle and Suzy's stone hits the bottle first and therefore smaches it, then we have a continuous line of causal processes and interactions from Suzy's throw, in which Suzy's arm exchanges energy with the stone, to the stone possessing kinetic energy while flying through the air, to the smashing of the bottle, in which the very same stone exchanges energy with the bottle. So, according to the naive process theory, Suzy's throw clearly is a causal producer of the smashing of the bottle. Whether Billy's stone would have smashed the bottle, in case Suzy would not have thrown, does not make any difference.

Dowe is convinced that NPT is a necessary condition for causal production, which would mean that every case of causal production involves a continuous line of causal processes and interactions as described in NPT. The reason that he calls it the *naive* process theory is that NPT is clearly not sufficient for causal production. Consider the following example by Dowe:

[10] The differentiation between causal processes and pseudo processes is a crucial aim of Salmon's and Dowe's process theories. See (Hitchcock, 2009b) for a criticism of both theories with regard to that aim.

[11] See (Dowe, 2000, p. 146). Notice that Dowe does generally not use the term 'causal production'. Instead, he typically simply speaks of 'causation' or of a 'causal relation' between events. But Dowe's examples clearly indicate that his process theory seeks to be an analysis of a production-conception of causation. In (Dowe, 2000, Chapter 6), Dowe calls cases of pure dependence-causation, like double-preventions or omissions of preventers, 'quasi-causation', and thereby explicitly separates pure dependence-causes from what he considers to be real causes.

> “Take any two causally independent events, say, my tapping the table and the clock hand moving a moment later. There is no causal connection between these two particular events, but there is in fact a set of causal processes and interactions between them, according to the Conserved Quantity theory. The reason is that there are air molecules filling the gap between them, so that one can connect the two events by stringing together a series of molecular collisions. Also, there is a longitudinal disturbance, the sound of my tapping, which reaches the clock hand before it moves” (Dowe, 2000, p. 149).

In a world that is overcrowded with attracting, repelling, and colliding molecules and that is traversed by all kinds of radiation, being connected by a continuous line of causal processes and causal interactions is rather simple. For any event E we can rifle through its past light cone and find arbitrarily many events that are connected to E by a continuous line of causal processes and causal interactions, even if most of those events a completely causally irrelevant to E.[12] Coming back to Sophie and Billy, this means that NPT also counts Billy’s throw as a cause of the smashing of the bottle. Colliding molecules, emitted or reflected photons, even the gravitation from Billy’s mass ensure that there is a continuous line of causal processes and causal interactions between Billy’s throw and the smashing of the bottle, even if his stone does not actually hit the bottle. Dowe calls cases like these, in which there is a continuous line of causal processes and causal interactions between C and E, even though it is intuitively clear that C does not causally produce E, *misconnections*.

One might hope for a simple solution to the misconnection-problem. We already know a condition that is sufficient for differentiating causally relevant from causally irrelevant factors: de facto dependence. So, if Dowe is right and NPT is a necessary condition for causal production, can’t we just combine NPT with a condition of de facto dependence to yield a necessary and sufficient condition for causal production? Can we say that a dependence-cause C of E, that additionally fulfills NPT, is a causal producer of E? The answer is no. Even if we use de facto dependence to preselect all events that are causally relevant for E, NPT is not strong enough to differentiate causal producers (or co-producers) from pure dependence-causes like double-preventers. As an example, consider our basketball scenario. Notice first, that it is no problem for NPT to recognize Jordan’s shot as a causal producer of the scoring. Roughly put, we have the following processes and interactions: Jordan’s shot amounts to an exchange of energy between his body and the ball, which then possesses a certain amount of kinetic energy until its world line intersects with the world line of the basket, where again an exchange of energy takes place that amounts to the scoring. But Pippen’s crossing of Johnson’s path to Jordan also satisfies NPT concerning the scoring. There are photons reflected from Pippen, that possess energy and ultimately interact with the basket during the scoring and if, for example, Pippen makes a noise, be it a little yelp after Johnson nearly bumped into him, there are sound waves, which are successions of colliding air molecules that proceed from Pippen to the basket. So, even in combination with a condition of de facto dependence, NPT is not sufficient for causal production. Dowe (2000, cf. p. 150 ff.) also considers several possibilities of combining NPT with some probabilistic or counterfactual account of causation, but he rejects them all, since none

[12] Hitchcock (1995) has made the same argument against Salmon’s process theory.

of these ‘integrating solutions’ succeeds in yielding a sufficient condition for causal production.

Dowe ultimately reaches for another solution to the misconnection-problem, which I will call the elaborated process theory (EPT) of causal production. For this, Dowe heavily restricts what kind of things can be the relata of causal production. He only allows physical facts or events that either have the form ‘object a has the amount x of conserved quantity q at time t’, in short: $q(a) = x$ at t, or the form ‘object a changes its amount of conserved quantity q from x to x' at t’. Admitting negative or disjunctive facts of the form ‘$q(a) \neq x$ at t’ or ‘$q(a) > x$ at t’ as causal relata would, as Dowe (2000, cf. p. 174-5) himself points out, lead to counterexamples against EPT. Dowe therefore explicitly excludes such multiply realizable events as causal relata. Dowe also constrains which continuous link of causal processes and causal interactions amounts to a causal connection between two physical events. He stipulates the following conditions for causal production:[13]

Connection of Causal Production. *There is a connection (or thread) of causal production between a fact $q(a) = x$ at t and a fact $q'(b) = y$ at t' if and only if there is a set of causal processes and interactions between $q(a) = x$ at t and $q'(b) = y$ at t' such that:*

(1) any change of object from a to b and any change of conserved quantity from q to q' occur at a causal interaction involving the following changes: $\Delta q(a)$, $\Delta q(b)$, $\Delta q'(a)$, and $\Delta q'(b)$; and

(2) for any exchange in (1) involving more than one conserved quantity, the changes in quantities are governed by a single law of nature.

Dowe (2000, cf. p. 172) then aims to extend the scope of the causal production relation by the following stipulation: There is a connection of causal production between the higher level events C_m and E_m only if C_m supervenes on the physical event C_p, E_m supervenes on the physical event E_p and there is a connection of causal production between C_m and E_m. Notice, though, that this extension of the scope of the causal production relation is rather limited, since disjunctive or multiply realizable events or facts are generally excluded as causal relata in Dowe’s theory.

EPT seems to be able to deal with some misconnection-cases that would have troubled NPT. Consider, for example, a steel ball a with linear momentum $\vec{x}$ ($q(a) = \vec{x}$), which is on course of colliding with a wooden ball b. On a’s way to b we electrically charge a without thereby changing its linear momentum $\vec{x}$. After the collision with b, in which a and b exchange linear momentum (but no charge), a has a new linear momentum $\vec{x}'$. Since there is a continuous link of causal processes and causal interactions from a’s charging to the physical event $q(a) = \vec{x}'$, NPT considers a’s charging as a causal producer of $q(a) = \vec{x}'$. But according to EPT, there is no connection of causal production between a’s charging and $q(a) = \vec{x}'$, because there is no interaction, in which a changes both its charge and its linear momentum in accordance with a single law of nature.[14] Anyhow, Hausmann (2002) and Ehring (2003) argue that EPT still falls prey to misconnection examples, which would mean that, just like NPT, EPT is not sufficient for causal production. Dowe (2018) picks up some examples from Hausmann and Ehring and

[13] See (Dowe, 2000, p. 171-2). I have slightly amended the notation.

[14] See (Dowe, 2018) for similar examples.

argues that his account can indeed handle them adequately. Be it as it may. I do not want to enter this debate here. The reason is this: even if EPT, or a slightly modified version of it, does not fall prey to misconnection examples, it still fails as an analysis of the concept of causal production, since it does not adequately describe what we typically mean by causal production and how we actually employ this concept.

Notice first that the explicit exclusion of disjunctive or multiply realizable events as causal relata makes Dowe's account of causal production much too narrow. As a consequence, a huge amount of prototypical examples of causal production cannot be captured by it. The shattering of a bottle, the outbreak of a forestfire, or having headache are all multiply realizable events. Dowe's account cannot tell us, what it means for such an event to be causally produced. But let us assume that there is still a good chunk of non-disjunctive higher-level events, to which Dowe's account of causal production can be applied. According to EPT, two such higher level events only stand in a relation of causal production, if the physical facts on which they supervene stand in a relation of causal production. So, whenever we want to evaluate whether two events stand in a relation of causal production to each other, we first have to identify their respective physical supervenience bases and evaluate, whether those supervenience bases stand in a relation of causal production to each other. But this is clearly not how we operate, when we assess relations of causal production. We typically feel able to tell, whether an event C causally produces another event E, even if we have no idea what the simple physical facts are, on which C and E supervene, or if they even supervene on physical facts at all. When Suzy insults Billy and thereby makes him sad, we have a clear example of causal production, even though it is entirely unclear, on what physical facts Suzy's insult and Billy's sadness supervene or if they even supervene on physical facts at all.[15]

Now, Dowe never claimed to provide a conceptual analysis, that accurately describes what we mean by causal production and how we actually employ this concept. Instead, Dowe (2000) is very clear that what he aims to achieve with EPT is an empirical analysis that "seeks to establish what causation in fact *is* in the actual world" (Dowe, 2000, p. 5. *Emphasis in original*). My point is not that EPT is false as an empirical analysis of causal production in our actual world. It may very well be the case, that in our actual world every relation of causal production is ultimately grounded in a relation between two physical facts that is adequately described by EPT. Instead, my point is this: Even if EPT is an adequate empirical analysis of causal production, it is still inadequate as a conceptual analysis if we employ the concept of causal production. Even if every relation of causal production is ultimately grounded in a certain causal relation between two physical facts, we typically feel confident to judge about the higher-level relation of causal production without having any knowledge about the underlying physical facts. This suggests, that the conditions, that we adhere to in our employment of the concept of causal production, do not make any reference to underlying physical facts and their respective causal relations. EPT, just like any other process-account of causation, is therefore unable to provide us with

[15] Another challenge for EPT is to explicate what a single law of nature is. Although Dowe states that "[n]o deep account of laws is required here - simple covariance will suffice" (Dowe, 2000, p. 172), we still need a clear cut criterion for deciding whether something is a single law of nature or the conjunction of several laws of nature. But here again, even if this challenge can be mastered, it does not prevent EPT's failure as an analysis of the concept of causal production.

an account that adequately describes how we actually use the concept of causal production and that adequately grasps its meaning.

6.3.2 The Mechanistic Account

There is another account of causal production that can be seen as an elaboration of the locality thesis: the mechanistic theory of causation by Stuart Glennan. Broken down, the theory claims that the *biff*, which connects a causal producer with its effect, is a mechanism. Here is Glennan's latest explication of this thesis:[16]

MC. *A statement of the form 'Event C causally produces event E' will be true just in case there exits a mechanims by which C contributes to the production of E.*

To grasp MC, we have to elaborate on two things. First, what exactly is a mechanism and second, how can a mechanism contribute to the production of an event? First things first, here is Glennan's definition of a mechanism:[17]

Mechanism. *A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon.*

The crucial constituents of a mechanism are entities and activities. Interactions are just activities that involve several entities. The entities that can serve in a mechanism are "fully determinate particulars located somewhere in space and time" (Glennan, 2017, p. 20) and they are "things that have reasonably stable properties and bounderies" (Glennan, 2017, p. 20). This encompasses objects like electrons, protons and atoms, but also higher-level objects like "proteins, organisms, congressional committees, and planets" (Glennan, 2017, p. 20). This holds similarly for activities, which can include "anything from walking to pushing to bonding (chemically or romantically)" (Glennan, 2017, p. 20). As Glennan points out, "[t]here cannot be activities without entities or entities without activities" (Glennan, 2017, p. 20). Now, crucially: "Activities produce change in entities (parts, components) that act or are acted upon" (Glennan, 2017, p. 31). The producing, that Glennan refers to, is understood to be causal. So, a mechanism is a collection of entities that, through their organized interrelations and their ability to act upon each other, form a cascade or a sequence of causal productions. The activities of the first group of entities produce a change in the activities of a second group of entities, which in turn produce a change in another group of entities, and so on. This is why Machamer et al. (2000) emphasize the existence of start-up and termination conditions of a mechanism. Once the start-up conditions are triggered, they will kick off a cascade of causal interactions that,

[16] See (Glennan, 2017, p. 156). Just like Dowe, Glennan does not use the term 'causal production'. Instead, he also generally talks about 'causation'. But his discussions in (Glennan, 2017, Chapter 6) and (Glennan, 2017, Chapter 7) make sufficiently clear, that MC is understood to capture the notion of causal production. Just like Hall (2004), Glennan differentiates between production and dependence, which he calls 'relevance'. He explains: "While I grant that production and relevance are two different concepts of cause, I will argue that production is fundamental. Both production and relevance claims are made true by objective features of mechanisms that underlie causal relationships, but I will argue that causal relevance claims are ultimately comparative claims about actual or possible productive causal mechanisms" (Glennan, 2017, p. 155). He also points out that omissions, though potentially causally relevant to an effect, "cannot contribute to the production of anything" (Glennan, 2017, p. 160). He later (Glennan, 2017, cf. p. 200) extends the same claim to (double-) preventers.

[17] See (Glennan, 2017, p. 17).

up until the termination conditions are reached, form the mechanism. In the present section, I follow the convention of using upper case roman letters to denote entities and upper case greek letters for activities. For an entity X engaging in the activity Φ, I will write $\Phi(X)$.

Notice that according to Glennan's definition, there is never a mechanism per se, but only a mechanism for a certain phenomenon, for which the mechanism is responsible. It is this relativization to a phenomenon that determines the bounderies of a mechanism. This phenomenon in question is itself some entity S engaging in a certain activity Ψ and the relationship between the mechanism and the phenomenon, for which it is responsible, is that of constitution. Figure 6.2 shows a so-called Craver Diagram that is supposed to illustrate a mechanism for the phenomenon $\Psi(S)$. The mechanism consists of those components or parts of S that, through their activities and interactions, constitute the activity Ψ. This does not necessarily involve all mereological components or parts of S. It may well be the case that the activities of only a few components of S are relevant for constituting a certain activity of S.[18]

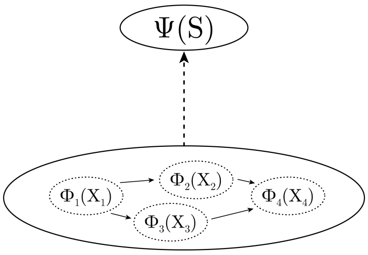


Figure 6.2: Craver Diagram - a mechanism constitutes activity Ψ of entity S. Adapted from (Craver, 2007, p. 7).

We are now already confronted with the first way by which a mechanism can produce an event. An event is simply some entity S engaging in an activity Ψ and a mechanism can produce $\Psi(S)$ by constituting it. Glennan calls this *constitutive production*. But there is another way by which a mechanism can contribute to the production of an event $\Psi(S)$. A mechanism can contain events, certain entities engaging in certain activities, that causally produce the event $\Psi(S)$ or (some of) the start up conditions of a mechanism that constitutes $\Psi(S)$. A prototypical case of causal production according to Glennan's account is illustrated in figure 6.3.

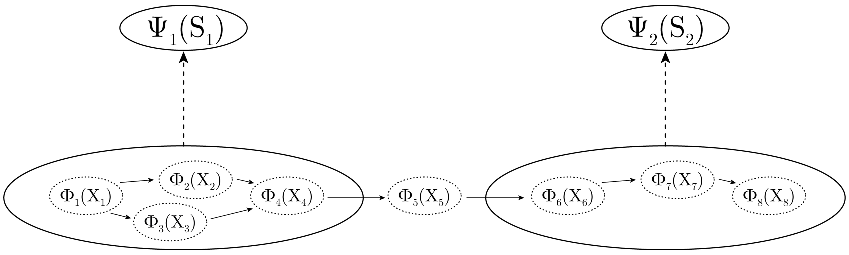


Figure 6.3: $\Psi_1(S_1)$ causally produces $\Psi_2(S_2)$.

Event $\Psi_1(S_1)$ causally produces the event $\Psi_2(S_2)$, because there is a mechanism constituting $\Psi_1(S_1)$ that contains events which causally produce the start-up conditions of the mechanism that constitutes $\Psi_2(S_2)$.

[18] See (Craver, 2007) or (Baumgartner and Gebharter, 2016) for conditions of constitutive relevance, which aim to enable the determination of just the components of an entity S that are responsible for the activity of S.

A first obvious criticism of Glennan's formulation of MC is that it does clearly not provide us with a sufficient condition of causal production. An event $\Phi(X)$ that is part of a mechanism that constitutes $\Psi(S)$ satisfies the condition in MC relative to $\Psi(S)$, since there is a mechanism by which $\Phi(X)$ contributes to the production of $\Psi(S)$. But $\Phi(X)$ does not causally produce $\Psi(S)$. Another criticism of MC may be directed at the vagueness of the term "contributes to the production". Glennan claims that a double preventer is no causal producer. But it is not clear why a double-preventer like Pippen's crossing of Johnson's path to Jordan is not constituted by a mechanism that contributes to the production of the scoring. One could easily claim that it contributes to the production of the scoring, because it prevents an event that would otherwise have prevented it. Both criticisms show that the exact formulation of MC is still in need of adjustments. But let us assume, that these difficulties can easily be dealt with.[19] The real danger for Glennan's account is something else: Circularity. If Glennan analyzes a given causal production by some underlying mechanisms, and mechanisms consist of events that causally produce other events, then Glennan clearly analyzes causal production in terms of causal productions.

Glennan is, of course, well aware of the circularity of his account. But he contends that the circularity is not vicious. When it comes to the analysis of a given token causal production, Glennan is right. His account does not analyze the causal production between two token events $\Psi_1(S_1)$ and $\Psi_2(S_2)$ in terms of the very same causal production between $\Psi_1(S_1)$ and $\Psi_2(S_2)$. Instead, his account analyzes the causal production between $\Psi_1(S_1)$ and $\Psi_2(S_2)$ in terms of several other causal productions that occur in the mechanisms that constitute the events $\Psi_1(S_1)$ and $\Psi_2(S_2)$. This is clearly not viciously circular. Quite the contrary, such an analysis of a given causal production can yield gainful insights, since it elaborates on a more detailed level of description how exactly the causal production proceeds.[20] But, and this is crucial, as a conceptual analysis of the concept of causal production, Glennan's account is indeed viciously circular. If we want to know the meaning of the concept of causal production and the general conditions under which we consider it to be justified to apply this concept, then we cannot be satisfied with conditions that already presuppose the correct application of the very same concept.

6.4 Sketching A New Account of Causal Production

6.4.1 In Search of a Distinguishing Feature of Causal Production

In accordance with (Hall, 2004), we have acknowledged that the concept of causal production, a concept of token causation that lies at the heart of our expanded Power PC theory, differs intensionally and extensionally from pure dependence-concepts of causation that are based on

[19] We could, for example, explicitly state that an event can only contribute to the production of $\Psi(S)$ by causally producing (some of) the start-up conditions of a mechanism that constitutes $\Psi(S)$ or by being constituted by a mechanism that contains events that causally produce $\Psi(S)$ or (some of) the start-up conditions of a mechanism that constitutes $\Psi(S)$.)

[20] Clearly, we might hit rock-bottom at some point and be confronted with events that have no underlying mechanisms. In this case, Glennan's account cannot be used as an analysis of the causal production between these two events. This is why Glennan explicitly considered his account to be an analysis of higher-level causal production.

the idea of a de facto counterfactual dependence of the effect on the cause. But while the framework of causal models enables us to formulate clear cut conditions to identify dependence-causes like actual, sufficient, or strong actual causes in concrete scenarios, we have come to notice that we lack clear cut criteria to identify what we intuitively recognize as causal producers and to differentiate such causal producers from mere dependence-causes like omissions of preventers or double preventers.

Hall (2004) emphasizes three features that production-causes generally possess and mere dependence-causes, at least in general, do not possess: transitivity, intrinsicness, and locality. As expressed in axiom C6 of the expanded Power PC theory, I agree with Hall that transitivity is a general feature of causal production. But transitivity does not provide a general distinguishing feature that helps us to differentiate causal producers from mere dependence-causes, because even though it is no general feature of dependence-causation, most mere dependence-causes still satisfy transitivity. Intrinsicness, though only attributed to causal producers, can also not help us to differentiate causal producers from mere dependence-causes, because the assumption of being a causal producer is the reason why we attibute an intrinsic causal power to it. We can therefore not use the attribute of having an intrinsic causal power as the distinctive feature by which we recognize causal producers in the first place. In the last section, I have argued, that it depends on the exact explication of locality, whether it is really a general feature of production-causation. If we understand locality in terms of the naive process theory (NPT), that is, as saying that there is a continuous line of causal processes and causal interactions between the cause and the effect, then locality clearly seems to be a general feature of production-causation. But it cannot serve as a distinguishing feature that helps us to differentiate causal producers from mere dependence-causes, because, as misconnection-examples show, most, or even all mere dependence-causes satisfy locality as well. If we, on the other hand, understand locality in terms of Dowe's elaborated process theory (ECT), then it remains rather uncertain, whether locality is actually a general feature of causal producers, since there are many intuitive cases of causal production, where it is entirely unclear whether there really is a connection between cause and effect that conforms with Dowe's criteria. So here again, locality is no generally reliable and epistemically accessible distinguishing feature that helps us to differentiate causal producers from mere dependence-causes. Understanding locality in terms of Glennan's mechanistic account (MC) is of no help either. To assess, whether something is a causal producer of a certain effect, we already need to know about the presence of other relations of causal production. So, as long as we lack general criteria that tell us, under what conditions a cause counts as a causal producer of a given effect, Glennan's explication of the locality thesis cannot help us to formulate such general criteria either.

What conclusions should we draw from this? Does this mean that there is simply no general distinguishing feature that helps us to differentiate causal producers from mere dependence-causes? Is there no other choice than understanding causal production as a basic, unanalyzable term? I do not think, that this is necessaritly true. In the present and the following two sections, I will lay out a path, that may ultimately lead us to a criterion which is able to differentiate causal producers from mere dependence-causes without relying on any kind of locality-thesis. The path, that I will lay out, will not amount to a full-fledged account of causal production.

This would simply go beyond the scope of the present dissertation. It can rather be seen as a first incomplete sketch of such an account. But this sketch will already provide us with a much clearer picture of causal production and prevention than we currently have and it will ultimately benefit the main goal of this dissertation, which is an enhanced understanding of causal explanations.

6.4.2 Solitary De Facto Difference Making

Any event C that causally produces a given effect E is a de facto difference maker to that effect. As long as there are no alternative or backup producers of E, the removal of C by an intervention would lead to a removal of E and the re-introduction of C by an intervention would lead to a re-emergence of E. In our expanded Power PC theory, this is ensured by the axioms C2: When C produces E, C is the case and E is the case, and C3: Any occurring event E is causally produced. So, causal producers of E are, just like omissions of preventers of E or preventions of preventers of E, dependence-causes of E.[21] But the question is, are there any characteristics of causal producers that distinguish them from mere dependence-causes, which are dependence-causes that are no causal producers?

I have already mentioned a widespread intuition that suggests an affirmative answer to this question, namely the intuition, that causal production is the more basic or essential kind of causation, while mere dependence-causes are rather derivative. According to this intuition, mere dependence-causes, like omissions of preventers or preventions of preventers, are only able to make a de facto difference to an effect E, because of the presence of a certain causal producer of E. Take our basketball scenario as an example. The omission of Johnson's block is only a difference maker to the scoring, because Jordan produced the scoring by throwing the ball in the first place. If Jordan would not have thrown, it would have been completely irrelevant whether Johnson did or did not block. Intuitively, the difference making ability of mere dependence-causes of an effect E therefore somehow piggy-back on the presence of certain causal producers of E. This is why Hitchcock (2007, cf. p. 496) calls cases, in which an effect counterfactually depends on an omission or a prevention of a preventer, 'parasitic dependence'. This parasitic character is also reflected in the fact, that the omitted or prevented preventer stands in a three-place relation to the effect E: It prevents a certain producer C from producing E. The producer C, on the other hand, stands in a 'direct' two-place causal relationship to the effect E. It therefore appears as a basic, non-parasitic causal relation.

One might hope that this intuition makes the difference between causal producers and mere dependence-causes of an effect E formally graspable. But a closer look reveals that this is not the case. It is true that, if C is a causal producer of E, then C's ability to produce E does not depend on the presence of any other causal producer of E. But it does depend on the absence of C-specific preventions of E. If A is a C-specific preventer of E that, if it would be the case, would

[21] Some authors have pointed out that there are some causal producers that are no dependence-causes. Glennan, for example, writes: "[I]n cases of overdetermination, where more than one cause may be sufficient for an effect, it appears that there is production without relevance" (Glennan, 2017, p. 154). But notice that in claims like these, relevance, which is basically Glennan's term for dependence-causation, is understood in terms of simple counterfactual dependence. As soon as we understand dependence-causation in terms of a de facto dependence, then there is no producer of a given effect, that is not also a dependence-cause of that effect.

prevent C's production of E in the given situation, then C's ability to be a de facto difference maker to E depends on the omission of A or the prevention of A's prevention of E. This shows that, rather than being parasitic, the relationship between mere dependence-causes and causal producers is better described as being a relation of mutual assistance. The dependence clearly goes both ways.

We have nonetheless gained an important insight. The mutual dependence of certain causes of an effect E basically yields a parcelling of E's causes into cohering packets. While the de facto difference making ability of any event inside of a such a packet depends on the presence of some other event in that packet, the de facto difference making ability of the packet as a whole does not depend on any other event. I will call the conjunction of all the events contained in such a packet a *solitary de facto difference maker for* E. As pointed out, mere dependence-causes, like omissions or preventions of preventers of E cannot be solitary de facto difference makers for E on their own, because they always rely on the presence of a certain causal producer of E that successfully produces E. This suggests the following preliminary hypothesis:[22]

Hypothesis 1 (Preliminary). *Every solitary de facto difference maker for* E *contains exactly one causal producer of* E *and any causal producer of* E *is part of a solitary de facto difference maker for* E*.*

A solitary de facto diffference maker of an effect E can very well consist only of a causal producer C of E. This is the case, if there are no C-specific preventers of E that, if they were the case in the given situation, would prevent C from producing E. But as soon, as there are C-specific preventers of E that, if they were the case in the given situation, would prevent C from producing E, the solitary de facto difference maker of E that contains C must also contain mere dependence-causes of E, namely the omissions or preventions of the respective C-specific preventers of E.

6.4.3 From Solitary De Facto Difference Making to Strong Actual Causation

Is there a way to formally explicate the concept of a solitary de facto difference maker? The answer is yes. And interestingly, we have already done so, namely with the definition of strong actual causation in the causal model framework. The conditions for strong actual causation basically demand the following: No matter what we do with the values of the other variables in the model that do not lie on an active causal path from the strong actual cause $\vec{X} = \vec{x}$ to the effect ϕ, $\vec{X} = \vec{x}$ still remains a de facto difference maker for ϕ. This means that the ability of $\vec{X} = \vec{x}$ to be a de facto difference maker for ϕ does not depend on any other factors in the model than those contained in $\vec{X} = \vec{x}$ itself, which is just the idea of a solitary de facto difference maker. This leads us to the following, again preliminary, hypothesis:

Hypothesis 2 (Preliminary). *Every strong actual cause of an event* ϕ *in a causal setting* $(\mathcal{M}, \vec{u})$ *contains exactly one causal producer of* ϕ *and any causal producer of* ϕ *in* $(\mathcal{M}, \vec{u})$ *is part of a strong actual cause of* ϕ *in* $(\mathcal{M}, \vec{u})$*.*

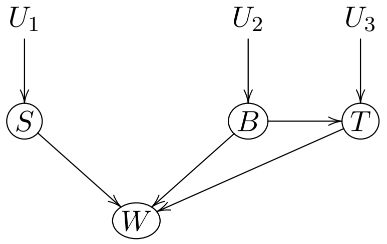
[22] We will later see, that this hypothesis needs a slight correction.

When we explicate the idea of solitary de facto difference making in terms of strong actual causation, we have to keep in mind, though, that the concept of strong actual causation relativizes the concept of solitary de facto difference making to a given causal setting $(\mathcal{M}, \vec{u})$. This amounts to a weakening of the model-independent concept of solitary de facto difference making that is used in hypothesis 1: The ability of an event $\vec{X} = \vec{x}$ to be a de facto difference maker for ϕ can be independent of all other factors that are represented in the causal setting $(\mathcal{M}, \vec{u})$, which means that $\vec{X} = \vec{x}$ is a strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$, but there may still be some factors that are not represented in $\mathcal{M}$, on which the de facto difference making ability of $\vec{X} = \vec{x}$ for ϕ depends. That would mean that $\vec{X} = \vec{x}$ is, from a global perspective, no solitary de facto difference maker for ϕ. This clearly calls the adequacy of hypothesis 2 in question. If ϕ has a causal producer that is not explicitly represented in the causal model $(\mathcal{M}, \vec{u})$, then ϕ may well have a strong actual cause in $(\mathcal{M}, \vec{u})$, that does not entail a causal producer of ϕ, but only mere dependence-causes like omissions or preventions of preventers.

This seems to suggest that, if we want to find the causal producers of a given effect ϕ, we should not employ the model-relative notion of solitary de facto difference making in form of strong actual causation, but that we should instead stick to the global, model-independent concept of solitary de facto difference making. But there are good reasons to believe that this would be a rashed judgement. First, the model-relative notion of solitary de facto difference making has a crucial advantage over a global, model-independent notion of solitary de facto difference making. While identifying a global, model-independent solitary de facto difference maker for a given effect ϕ is in theory a nice ideal to strive for, it is practically nearly impossible to do so. Causal reasoning always works on the basis of simplifying models of real-life scenarios and these models typically abstract away from several background conditions that are assumed to remain sufficiently constant. A model-relative concept of solitary de facto difference making is therefore the only viable option, if we aim to apply the concept on real-life scenarios. But there is more than just a pragmatic advantage of the model-relative concept of solitary de facto difference making. The model-relative concept of solitary de facto difference making in terms of strong actual causation even seems to yield a superior description of how we actually use the concept of causal production than the global, model-independent concept of solitary de facto difference making. The global, model-independent concept of solitary de facto difference making leads to a global, model-independent concept of causal production. The model-relative concept of solitary de facto difference making in terms of strong actual causation, on the other hand, leads to a model-relative concept of causal production. As we will later see, it is this model-relative version that actually conforms better with our intuitive understanding of causal production.

6.4.4 Default and Deviant States

If we aim to identify the causal producers of a given effect, then, according to hypothesis 2, employing the concept of strong actual causation forms a solid initial step. To see why, consider the following example: Suzy throws a baseball at a window ($S = 1$) and so does Billy ($B = 1$). Both throw and hit the window at the same time, but each baseball on its own would have been sufficient to break the window ($W = 1$). Somewhere close to Billy stands Tom, Billy's big brother, who normaly tries to keep him out of trouble. He thought about jumping in front of Billy to catch his ball out of mid air ($T = 1$). But for some reason he decided not to do so ($T = 0$). We can represent the scenario by the causal model $\mathcal{M}^{Ba}$ as shown in figure 6.4.



- $S := U_1$
- $B := U_2$
- $T := B \wedge U_3$
- $W := S \vee (B \wedge \neg T)$

Figure 6.4: $\mathcal{M}^{Ba}$ - the baseball scenario.

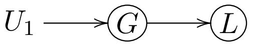
In the given situation, we have $U_1 = U_2 = 1$ and $U_3 = 0$. If we apply the definition of strong actual causation, we get that $S = 1$ is a strong actual cause of $W = 1$ and $B = 1 \wedge T = 0$ is a strong actual cause of $W = 1$.[23] Intuitively, $S = 1$ and $B = 1$ are both causal producers of $W = 1$, while $T = 0$ is not. This is in accordance with our hypothesis 2. Since we are facing a case of overdetermination, we have two causal producers of the effect $W = 1$. The concept of strong actual causation keeps distinct causal producers apart: each causal producer appears in a separate strong actual cause of the effect. $B = 1 \wedge T = 0$ is an example of a strong actual cause that contains more than just a causal producer. This illustrates that the concept of strong actual causation is unable to further differentiate between a causal producer ($B = 1$) and the omission of a preventer ($T = 0$). We can therefore think of the concept of strong actual causation as a sieve that is still somewhat too coarse for our target, namely the concept of causal production. Applying it will give us the causal producers of a given effect ϕ. It will also separate distinct causal producers of ϕ in cases of overdetermination by parcelling them in distinct packets. But, in general, it gives us each causal producer $\vec{C} = \vec{c}$ of ϕ in a bundled packet that, besides the producer, may also contain mere dependence-causes, like omissions or preventions of ($\vec{C} = \vec{c}$)-specific preventers of ϕ.

I have already indicated that hypothesis 2 is only preliminary. To see why, consider the following example. According to the Bible, the following happend:

> "In the beginning God created the heaven and the earth. And the earth was without form, and void; and darkness was upon the face of the deep. [...]. And God said, Let there be light: and there was light" (The Bible, 1998, Genesis 1. 1-3).

[23] Here and in the following examples I will not explicitly state the proofs for claims about relations of strong actual causation. The claims can easily be verified by checking the fulfillment of the conditions provided in the definition of strong actual causation as presented in chapter 1.

We can use the rather simple causal model $\mathcal{M}^G$ as shown in figure 6.5 to represent this situation.



- $G := U_1$
- $L := G$

Figure 6.5: $\mathcal{M}^G$ - the Genesis scenario.

Both variables are bivalent. $G = 1$ represents that God says “Let there be light” and $G = 0$ represents that God says nothing.[24] $L = 1$ represents that there is light, while $L = 0$ represents that darkness is upon the face of the deep. We now look at the actual situation, in which $U_1 = 1$. When applying the definition of strong actual causation, we get that $G = 1$ is a strong actual cause of $L = 1$ in the given causal setting and I consider it as intuitively adequate to say that God’s utterance “Let there be light” is a causal producer of there being light. So, in accordance with hypothesis 2, the only strong actual cause of $L = 1$ contains (and even is identical to) the only causal producer of $L = 1$. But now imagine that things went somewhat differently back then. God decided to remain silent, because she wanted to enjoy the darkness a bit longer. So, we have $G = 0$ and accordingly $L = 0$. Yet again, when applying the definition of strong actual causation, we get that $G = 0$ is a strong actual cause of $L = 0$. But it seems highly unintuitive to claim that God remaining silent causally produced the darkness. As mentioned shortly before in the Genesis, the darkness was already there. God’s silence is indeed a solitary difference maker for the darkness, because God could have said something and would thereby have created light. But God’s silence is no causal producer of the darkness, because it was the default state of the world, which does not need a causal production in the first place.

We have now an example that contradicts hypothesis 2 (as well as hypothesis 1). We have a strong actual cause (even a global, model-independent solitary difference maker) of an effect ϕ that does not entail a causal producer of ϕ. Instead, it only entails the omission of a causal producer of a complement of ϕ. The reason for the violation of hypothesis 2 (and hypothesis 1) seems to be this: The strong actual cause of ϕ does not entail a causal producer of ϕ, because ϕ does not have any causal producer in the first place. The event ϕ is considered to be a causally unproduced default state that remains to be the case as long as nothing causally produces an event that is incompatible with it.

Here is another, more mundane example that concerns our two vandals, Suzy and Billy. Take again the baseball scenario from above. Only this time, we have $U_1 = U_2 = U_3 = 0$, so Suzy and Billy both do not throw their balls, Tom does not jump in front of Billy and the window remains intact. When we now apply the definition of strong actual causation, we get that: $S = 0 \wedge B = 0$ is a strong actual cause of $W = 0$. But, intuively, neither $S = 0$, nor $B = 0$, nor their conjunction is a causal producer of $W = 0$. So, here again, we have a strong actual cause that does not contain a causal producer of the effect $W = 0$ and therefore another violation of hypothesis 2.

[24] Let us imagine that these are the only two alternatives for God.

How can we explain the violation this time? $W = 0$ does not seem to represent a causally unproduced default state, like the world being in darkness at the beginning of time. From a global, model-independent view, the window being intact ($W = 0$) is clearly the result of a causal prodution. It was produced in some glass factory, carefully delivered and built into the house that is now vandalised. It is therefore not a genuinely causally unproduced event per se. But it is a state that is assumed to be self-sustaining in the confines of the localized causal scenario that we are now considering. The window was intact before Billy and Suzy set their eyes on it and it is assumed to remain intact as long as there are no interferences that causally produce a change of its state. The fact that the window being intact was itself once causally produced does not change its status as a default state in the given context. What is important is only the assumption of it being a self-sustaining initial condition in the given localized causal scenario. Being a default state is therefore a model- or context-relative feature.

This context-relative understanding of default states is in accordance with both Maudlin's (2004) and Hitchcock's (2007) characterization of default states. According to Maudlin, the designation of default states depends on the acceptance of two kinds of laws: "*inertial* laws that describe how some entities behave when nothing acts on them" (Maudlin, 2004, p. 431. *Emphasis in original.*) and "laws of *deviation* that specify in what conditions, and in what ways, the behaviour will deviate from the inertial behaviour" (Maudlin, 2004, p. 431. *Emphasis in original*). He considers Newton's first and second law as a prototype of such a system of laws and therefore calls similar systems of inertial and deviation laws *quasi-Newtonian.* He further points out, that the designation of default states "depends on how we carve the situation up into systems. [...] Carving up the world differently can give different (special science) *laws*, governing different *systems*" (Maudlin, 2004, p. 436-7. *Emphasis in original.*), which amounts to a model-relativity of default states. Similarly, Hitchcock explains:

> "[T]here are certain states of a system that are self-sustaining, that will persist in the absence of any causes other than the presence of the state itself: the default assumption is that a system, once it is in such a state, will persist in such a state. Theory – either scientific or folk – informs us which states are self-sustaining" (Hitchcock, 2007, p. 506).

He then adds: "The default may depend upon the level of analysis" (Hitchcock, 2007, p. 506).

With the model-relative concept of a default state at hand, we can now explain, why hypothesis 2 is violated in the previous example: The strong actual cause of $W = 0$ does not contain any causal producer of $W = 0$, because $W = 0$ is considered to be a self-sustaining default state in the given localized causal scenario. This is why, in the confines of the given causal model, $W = 0$ is considered to be causally unproduced. This motivates the following reformulation of hypothesis 2:

Hypothesis 3. *Every strong actual cause of ϕ in a causal setting $(\mathcal{M}, \vec{u})$ contains exactly one causal producer of ϕ, as long as ϕ is no default state in $(\mathcal{M}, \vec{u})$, and if ϕ does have a causal producer in $(\mathcal{M}, \vec{u})$, then it is part of a strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$.*

Using the model-relative features of being a default-state and being a strong actual cause in hypothesis 3 yields a conception of causal production that is also model-relative. Embracing this model-relativity of causal production can explain some of our seemingly conflicting intuitive judgments about causal production. Consider the following example by Schaffer (2004). Chopping off a person's head is not only an obvious cause of the person's dead, it is intuitively also a causal producer of the person's death. Not chopping off the person's head, on the other hand, is intuitively not a causal producer of the person being alive. The reason for these intuitions is that we typically take being alive as a default state of a living organism. We typically take for granted all the intricate mechanisms that quietly take place in a person's body to keep it alive. They are simply assumed to continue as long as nothing extraordinary happens that interferes with them. Dying is therefore seen as the deviant state and, intuitively, it is always the deviant state that is understood to be in need of a causal producer.

As soon as we widen our perspective and consider that a person being alive is actually causally produced by the intricate and well aligned activities of the person's body, our intuitive assessment of the causal role of the beheading changes. It is under this widened consideration that we revisit our assessment of the default state. It is not being alive which is considered to be a default state anymore, but being dead. Following hypothesis 3, this leads to a reassessment of the causal production relation between the beheading and the dying. Since dying is now considered to be a default state, it is seen as causally unproduced. The beheading therefore now appears as a prevention of the body's activities from producing the person's vitality.

Here is another example by Schaffer (2004). Pulling the trigger of a loaded gun is intuitively a causal producer of the gun going off. Not pulling the trigger, on the other hand, is intuitively not a causal producer of the gun not going off. Here again, this is because we tyically consider that not going off is the default state of a gun, which is therefore understood as causally unproduced, while going off is typically seen as the deviant state of a gun, which is therefore understood as being in need of a causal production. But this intuition might change when having a closer look into the composition of the gun. There we might see that, as soon as a gun is loaded, a spring is compressed that by uncoiling produces the gun to fire. It is only the presence of an interposed seal that prevents the spring from uncoiling and therefore the gun from firing. Pulling the trigger removes the seal and thereby enables the spring to uncoil and the gun to fire. This new perspective, which takes the inner mechanism of a loaded gun into consideration, leads to a reasssessment of what we consider to be the default state of a loaded gun. We have learned that as soon as the gun is loaded, it would go off, were it not for the presence of a seal that keeps a certain spring from uncoiling. So, the firing of the gun is now seen as the default state, while the non-firing is the deviant state that needs a causal producer in the form of an interposing seal. Pulling the trigger now appears not as the producer of the firing, but as a prevention of the production of the non-firing.[25] Examples like these show: A change in perspective, a wider or more detailed analysis of the considered scenario, may change our assessments of default and deviant states and accordingly our intuitions about causal production. Hypothesis 3 and the associated acknowledgement of the model-relativity of the concept of causal production can explain, why intuitions about causal production may sway under changes in perspective.

[25] See (Maudlin, 2004) and (Hitchcock, 2007) for a similar analysis of Schaffer's examples. But neither Maudlin nor Hitchcock employs default values for an analysis of causal production as I propose it here. Maudlin aims at an analysis of causation, that is based on laws and that does not employ counterfactuals. Hitchcock, on the other hand, employs default values for an account of actual causation.

Based on hypothesis 3, we can now formulate some instructions that can be seen as the first steps of a procedure for identifying the causal producers of a given effect ϕ in a causal setting $(\mathcal{M}, \vec{u})$:

Step 1: Determine whether ϕ is a default or a deviant state in $(\mathcal{M}, \vec{u})$.

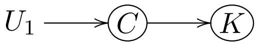
If ϕ is a default state in $(\mathcal{M}, \vec{u})$, then ϕ does not have any causal producer in $(\mathcal{M}, \vec{u})$. If ϕ is a deviant state in $(\mathcal{M}, \vec{u})$, then:

Step 2: Determine the strong actual causes of ϕ in $(\mathcal{M}, \vec{u})$.

Each strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$ will contain exactly one causal producer of ϕ in $(\mathcal{M}, \vec{u})$ and each causal producer of ϕ in $(\mathcal{M}, \vec{u})$ will be in a strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$.

The formal framework of causal models does not yet contain a classification of its variable-values into default and deviant states. The information about default and deviant states in a given causal setting is therefore external to any classical causal model. It would be possible, though, to expand the classical causal model framework by including a function $\mathfrak{d}$ that maps the values of each variable V in the model $\mathcal{M}$ on one of two values: either on the value *default* or on the value *deviant*. A causal setting of the form $(\mathcal{M}, \vec{u}, \mathfrak{d})$ then includes a complete attribution of default and deviant states to all its variable-values.

Notice that not every variable in a causal model must have a default value. Consider, for example, the following scenario: Before the start of a football game between Team 1 and Team 2, the referee throws a coin to decide who gets the kickoff. We can represent the scenario by a causal model as shown in figure 6.6, where the bivalent variable C represents the outcome of the coin-toss and the bivalent variable K represents which team gets the kickoff.



- $C := U_1$
- $K := C$

Figure 6.6: Causal model of the football scenario.

If the coin lands on heads ($C = 1$), then Team 1 gets the kickoff ($K = 1$). If the coin lands on tails ($C = 2$), then Team 2 gets the kickoff ($K = 2$). But neither $K = 1$, nor $K = 2$ is intuitively understood to be a default state. So according to hypothesis 3, a strong actual cause of $K = 1$ would contain a causal producer of $K = 1$, if $K = 1$ would be the case, just as well as a strong actual cause of $K = 2$ would contain a causal producer of $K = 2$, if $K = 2$ would be the case. This seems intuitively just right. If the coin lands on heads, it seems intuitively adequate to see this as a causal producer of Team 1 getting the kickoff. And if the coin lands on tails, it seems intuitively adequate to see this as a causal producer of Team 2 getting the kickoff.

We have now come some way. But our two-step instruction, which is based on hypothesis 3, is in general not yet able to uniquely identify the causal producers of a given effect ϕ in a causal setting $(\mathcal{M}, \vec{u})$. We have already seen that, besides a causal producer of ϕ, a strong actual cause of ϕ may also contain mere dependence-causes of ϕ. And so far we have no formal tools to differentiate the causal producer from the mere dependence-causes in a strong actual cause. So, the crucial question is: can we get any closer? Can we find a condition that helps us to differentiate causal producers from omissions or preventions of preventers in a complex strong actual cause? To answer this question, we first have to explicate what exactly prevention is.

6.5 Prevention

6.5.1 Dowe's Counterfactual Theory of Prevention

If the prevention of an event E is not the same as the production of $\neg E$, what else might prevention be? In our expanded Power PC theory we have assumed a crucial feature of prevention that contradicts PPC, namely that prevention is basically a three-place relation: There is no prevention of an effect E per se, but only preventions of certain productions of E. Interestingly, this idea accords well with Phil Dowe's counterfactual theory of prevention as he presents it in (Dowe, 2000, Chapter 6) and (Dowe, 2001). Although Dowe objects to a reduction of prevention to causal production in the spirit of PPC, due to the intuition of difference, he argues that prevention can be reduced to counterfactual claims about causal production. According to his theory, a claim about prevention "should be understood not as being directly about actual genuine causation but primarily as a counterfactual claim about genuine causation" (Dowe, 2001, p. 216). Building on his CQ theory of causal production, Dowe formulates the following conditions for token prevention:[26]

Prevention (Dowe, 2000). *A prevented E if*

(P1) A occured and E did not, and there occured an x such that

(P2) there is a causal relation between A and the process due to x, such that either

(i) A is a causal interaction with the causal process x, or

(ii) A causally produces y, a causal interaction with the causal process x.

(P3) if A had not occured, x would have causally produced E

Dowe considers the conditions P1-P3 to be sufficient for prevention. But he acknowledges that they are not necessary for prevention, due to the potential presence of other preventers besides A that could thwart the counterfactual dependence of the non-occurrence of E's production by x on the occurrence of A. Imagine that A did prevent x from producing E. If a back-up preventer D is present that would have stopped x's causal production of E, if A would not have stopped it, then A does not satisfy P3, even though it actually did prevent x from producing E. To deal with this problem, Dowe (2001, p. 134) proposes the following amendment of condition P3:

(P3') P3 or there exists a D such that had neither A nor D occured, x would have causally produced E or...

[26] See (Dowe, 2000, p. 132). I slightly amended the notation. As pointed out before, Dowe does not use the term 'causal production'. Instead, he uses the term 'causation' or 'genuine causation' for what I understand causal production to be.

So, if there is an event D that can also prevent x from producing E, the added disjunct in P3' makes sure that we check the counterfactual dependence of the non-occurrence of E's production by x on the occurrence of A not in the contingency, in which the alternative preventer D is present, but in the contingency, in which D is also considered to be absent. If there is more than one additional back-up preventer, Dowe simply proposes to amend P3' by adding more disjuncts, like '*...or there exists a D_1 and a D_2 such that...*'.

The crucial role of the causal process x in Dowe's definition of prevention clearly shows that Dowe incorporates the intuition that a preventer prevents an effect E first and foremost by preventing a causal producer from producing E. In addition to that, the following thesis can be seen as the core of Dowe's prevention-account:

(Prev). *A prevents x from producing E if and only if A is responsible for the fact that x does not produce E, while this responsibility is grounded in the following two facts:*

(a) The non-occurrence of E's production by x is de facto dependent on the occurrence of A.

(b) A causally produces some interaction with the causal process x.

I consider this core idea of Dowe's prevention-account to be highly intuitive. Nonetheless, I have my issues with several aspects of how Dowe explicates these core ideas in his definition displayed above.

First, although Dowe clearly incorporates the intuition that prevention is, first and foremost, a three-place relation, he still defines prevention as a two-place relation by simply demanding that *some* causal process x of E must exist such that A prevents x from producing E. But I consider the definition of prevention as a three-place relation as preferable, because, as shown in chapter 5, only cause-specific preventers have intrinsic preventive powers, which are an indespensible tool for causal reasoning. Secondly, Dowe builds his definition of prevention on a conception of causal production that is explicated in terms of his CQ theory. But, as argued in section 6.3, the CQ theory does not provide an adequate conceptual analysis of causal production. This is why I would like to avoid a CQ-based explication of causal production in an analysis of prevention. Thirdly, Dowe's explication of de facto dependence in terms of condition P3' does not work. Condition P3' actually introduces more problems than it solves. Consider a scenario, in which A prevents x's production of E. Let us further say that D is some event that also occurs and that has some influence on the causal process x but one that neither interrupts nor supports x's production of E. Although D is not able to prevent x's production of E, it does fulfill conditions P1, P2 and P3'. It fulfills P3', because if D and the actual preventer A would not have happened, x would have produced E. So, in general, as soon as we have a preventer of x's production of E, anything that has some non-essential causal influence on x counts as a preventer of x's production of E.

Here is a concrete example: Suzy drives a car on cruise control and heads towards a wall. If she would not brake, she would crash into the wall. Luckily, she brakes early enough. The braking causally interacts with a causal process x (the car having a certain amount of linear momentum) that would otherwise have led to the crash. The braking therefore prevents the crash. This is intuitively adequate. But imagine that Suzy does not only brake, she also sings her favorite song. Her singing produces sound waves and therefore leads to a causal interaction with the car and, accordingly, with the causal process x. According to Dowe's definition, this makes Suzy's singing a preventer of the crash: Suzy's singing occurs and the crash does not, so P1 is fulfilled. Suzy's singing has a causal interaction with the causal process that would have led to the crash, so P2 is fulfilled. And there is an event A, namely Suzy's braking, such that if A and Suzy's singing would not have occured, x would have led to the crash, so P3' is fulfilled.

I will now try to solve these issues while holding on to the core idea Prev of Dowe's counterfactual theory of prevention.

6.5.2 A New Counterfactual Theory of Prevention

The Problem of finding an adequate explication of *de facto dependence* is already familiar from the debates about the concept of actual causation. In this context, authors like Halpern and Pearl have shown that causal models provide a very suitable formal framework for explicating the relation of de facto dependence between a token effect and its actual cause. This suggests that the causal model framework can likewise be used to explicate the relation of de facto dependence that holds between the non-occurrence of an event's causal production and the occurrence of a preventer. But if we want to use the framework of causal models to formulate a definition of prevention, in which the concept of causal production plays a crucial role, we have to face the problem that we have not yet found criteria that can reliably identify the relation of causal production in causal models. For now, our only choice is therefore to treat causal production as an unanalyzed, basic concept and to ammend a classical causal setting $(\mathcal{M}, \vec{u})$ with an additional set *Prod* of event-tuples, such that $(\mathcal{M}, \vec{u}, Prod) \models (\vec{C} = \vec{c}$ produces $\vec{E} = \vec{e})$ if and only if $\vec{C} = \vec{c}$ is part of a strong actual cause of $\vec{E} = \vec{e}$ in $(\mathcal{M}, \vec{u})$ and $(\vec{C} = \vec{c}, \vec{E} = \vec{e}) \in Prod$. So, *Prod* basically entails the information, which causal relations in a causal setting amount to relations of causal production.

My final proposal for defining prevention in the framework of causal models will be rather complex. I will therefore slowly approach my final proposal by first considering two less complex preliminary versions. The shortcomings of these proposals and their required corrections will then ultimately lead to the final version and they will illustrate why its additional complexity is warranted. The first preliminary proposal to define the concept of prevention is formulated in a strict analogy to the HP-definition of actual causation:[27]

Actual Prevention (Proposal 1). *$\vec{X} = \vec{x}$ prevents $\vec{C} = \vec{c}$ from causally producing $\vec{E} = \vec{e}$ in the causal setting $(\mathcal{M}, \vec{u}, Prod)$ if the following conditions hold:*

Pr1 $(\mathcal{M}, \vec{u}, Prod) \models \vec{X} = \vec{x} \wedge \vec{C} = \vec{c} \wedge \neg(\vec{C} = \vec{c}$ *produces* $\vec{E} = \vec{e})$.

Pr2 *There is a partition $(\vec{Z}, \vec{W})$ of the endogenous Variables $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w})$ such that if $(\mathcal{M}, \vec{u}, Prod) \models Z = z^*$ for all $Z \in \vec{Z}$, then:*

(a) $(\mathcal{M}, \vec{u}, Prod) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}](\vec{C} = \vec{c}$ *produces* $\vec{E} = \vec{e})$.

(b) $(\mathcal{M}, \vec{u}, Prod) \models [\vec{X} \leftarrow \vec{x}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]\neg(\vec{C} = \vec{c}$ *produces* $\vec{E} = \vec{e})$ *for all $\vec{W}' \subseteq \vec{W}$ and for all $\vec{Z}' \subseteq \vec{Z}$.*

[27] Whenever I use the term 'prevention', I mean *actual* prevention, since what I aim to define here is the concept of token prevention in a concrete token situation.

Pr3 *$\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ (with $\vec{x}'$ being the restriction of $\vec{x}$ to the variables in $\vec{X}'$) satisfies conditions (Pr1) and (Pr2).*

The definition explicitly treats prevention as a three-place relation between the preventer $\vec{X} = \vec{x}$, the generative cause $\vec{C} = \vec{c}$, and the effect $\vec{E} = \vec{e}$. As a consequence, condition Pr1 deviates from Dowe's condition P1 in a crucial respect: While Dowe demands in P1 that $\vec{E} = \vec{e}$ must not be the case, Pr1 is actually compatible with $\vec{E} = \vec{e}$ being the case in the actual situation, because even if $\vec{X} = \vec{x}$ successfully prevents $\vec{C} = \vec{c}$ from producing $\vec{E} = \vec{e}$, it may still be the case that $\vec{E} = \vec{e}$ is produced by some other cause apart from $\vec{C} = \vec{c}$.

Condition Pr2 explicates the de facto dependence of the non-occurrence of the causal production of $\vec{E} = \vec{e}$ by $\vec{C} = \vec{c}$ on the occurrence of $\vec{X} = \vec{x}$. The success of condition Pr2 in capturing the idea of de facto dependence was the main motivation for employing the framework of SEMs for our definition of prevention. Reconsidering our previous example about Suzy's prevention of a car crash, illustrates that Pr2 does a much better job than Dowe's condition P3': In an adequate causal model of the scenario, Suzy's braking will still turn out as a prevention of the crash, since we have a simple counterfactual dependence between the non-occurrence of the crash and Suzy's braking. But Suzy's singing will not turn out as a prevention of the crash, because we do not have a de facto dependence of the non-occurrence of the crash on Suzy's singing in terms of Pr2, since Suzy's singing is unable to satisfy condition Pr2(b): It cannot ensure that the crash will not be produced by the moving vehicle in a contingency, in which Suzy does not brake. Finally, condition Pr3 is there to make sure that a preventive cause does not entail any irrelevant parts, which is another aspect that was missing in Dowe's prevention-account.

While Proposal 1 has already improved on many problematic aspects of Dowe's prevention-account, it still has some crucial flaws. First and foremost, it is missing an explication of Prev-(b): There is no condition which demands, that the preventer $\vec{X} = \vec{x}$ causally produces an interaction with the causal process by which $\vec{C} = \vec{c}$ would otherwise have produced $\vec{E} = \vec{e}$. Although Proposal 1 does indeed incorporate the idea that the preventer $\vec{X} = \vec{x}$ is responsible for the fact that $\vec{C} = \vec{c}$ does not produce $\vec{E} = \vec{e}$, this responsibility is only grounded in the de facto dependence of the non-occurrence of the production on the occurrence of $\vec{X} = \vec{x}$. But, as we will now see, this is not enough.

Consider again Suzy's prevention of the car crash. But imagine now, that Suzy is not alone. Billy sits on the passenger seat next to her. Yet again, Suzy brakes and thereby prevents the car from crashing into a wall. Billy does not do anything. Still, the fact that Billy did not stop Suzy from braking, for example, by silently sitting next to her, instead of distracting her with an interesting philosophical problem, also counts as a prevention of the crash according to Proposal 1. The non-occurrence of the production of the crash de facto depends on Billy abstaining from distracting Suzy. But it is quite strange to give Billy any credit as an actual preventer of the crash. Intuitively, he did not do anything that helped avoiding the crash. He only abstained from preventing Suzy from preventing the crash.

The example illustrates why a condition in the spirit of Prev-(b) is needed for a sufficient condition of actual prevention. A de facto difference maker for the non-occurrence of the production of the effect is not enough for what we intuitively call a preventer. The preventer must somehow be active. It must actively interfere with the causal production of the effect. This is what condition Prev-(b) aims to capture. But how can we explicate Prev-(b) in the framework of causal models without relying on concepts from Dowe's CQ theory, like *causal process* and *causal interaction*?

Here is my proposal: Typically, a causal producer $\vec{C} = \vec{c}$ of an effect $\vec{E} = \vec{e}$ does not produce $\vec{E} = \vec{e}$ directly, but via several causal mediators. In chapter 4, we have defined such a sequence of events, in which every event causally produces its direct successor, as an active path of production. We can now capture the idea of Prev-(b) by demanding that the preventer $\vec{X} = \vec{x}$ causally (co-)produces an event $\vec{Y} = \vec{y}$ such that, if $\vec{X} = \vec{x}$ would not have done so, the alternative $\vec{Y} = \vec{y'}$ would have been the case and $\vec{Y} = \vec{y'}$ would have been part of an active path of production that has $\vec{C} = \vec{c}$ as a predecessor and $\vec{E} = \vec{e}$ as a successor. I will call the variables $\vec{Y}$, the *variables of collision* (VoC) of the ($\vec{C} = \vec{c}$)-specific prevention of $\vec{E} = \vec{e}$ by $\vec{X} = \vec{x}$. This leads us to the following proposal:

Actual Prevention (Proposal 2). *$\vec{X} = \vec{x}$ prevents $\vec{C} = \vec{c}$ from producing $\vec{E} = \vec{e}$ in the causal setting $(\mathcal{M}, \vec{u}, Prod)$ if and only if:*

Pr1 $(\mathcal{M}, \vec{u}, Prod) \models \vec{X} = \vec{x} \wedge \vec{C} = \vec{c} \wedge \neg(\vec{C} = \vec{c}$ *produces* $\vec{E} = \vec{e})$.

Pr2 *There is an event $\vec{Y} = \vec{y}$ with $(\mathcal{M}, \vec{u}, Prod) \models \vec{X} = \vec{x}$ produces $\vec{Y} = \vec{y}$, such that there is a partition $(\vec{Z}, \vec{W})$ of the endogenous Variables $\mathcal{V}$ in $\mathcal{M}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w})$ such that if $(\mathcal{M}, \vec{u}, Prod) \models Z = z^*$ for all $Z \in \vec{Z}$, then:*

(a) $(\mathcal{M}, \vec{u}, Prod) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}](\vec{C} = \vec{c}$ *produces* $\vec{Y} = \vec{y'}) \wedge (\vec{Y} = \vec{y'}$ *(co-)produces* $\vec{E} = \vec{e})$.

(b) $(\mathcal{M}, \vec{u}, Prod) \models [\vec{X} \leftarrow \vec{x}, \vec{W'} \leftarrow \vec{w}, \vec{Z'} \leftarrow \vec{z^*}]\neg(\vec{C} = \vec{c}$ *produces* $\vec{E} = \vec{e}) \wedge (\vec{X} = \vec{x}$ *(co-)produces* $\vec{Y} = \vec{y})$ *for all* $\vec{W'} \subseteq \vec{W}$ *and for all* $\vec{Z'} \subseteq \vec{Z}$.

Pr3 *$\vec{X}$ is minimal; there is no strict subset $\vec{X'}$ of $\vec{X}$ such that $\vec{X'} = \vec{x'}$ (with $\vec{x'}$ being the restriction of $\vec{x}$ to the variables in $\vec{X'}$) satisfies conditions (Pr1) and (Pr2).*

Conditions Pr1 and Pr3 are the same as before. The crucial difference to Proposal 1 lies in condition Pr2. Condition Pr2 now incorporates formal explications of both Prev-(a) and Prev-(b). The non-occurrence of the production of $\vec{E} = \vec{e}$ by $\vec{C} = \vec{c}$ must be de facto dependent on the occurrence of $\vec{X} = \vec{x}$. But additionally, $\vec{X} = \vec{x}$ must causally (co-)produce an event $\vec{Y} = \vec{y}$ such that, if $\vec{X} = \vec{x}$ would not have done so, the alternative $\vec{Y} = \vec{y'}$ would have been the case and $\vec{Y} = \vec{y'}$ would have been part of an active path of production that has $\vec{C} = \vec{c}$ as a predecessor and $\vec{E} = \vec{e}$ as a successor. This means that $\vec{X} = \vec{x}$ causally (co-)produces an interruption of the process by which $\vec{C} = \vec{c}$ would otherwise have causally produced $\vec{E} = \vec{e}$.

Proposal 2 deals adequately with the scenario, in which Suzy brakes and thereby prevents a crash while Billy sits silently on the passenger seat. Billy's behaviour does not causally produce any interruption of the causal process by which the moving car would otherwise have causally produced the crash. Suzy, on the other hand, does. This is why her braking counts as an actual prevention of the crash, while Billy's restraint from disturbing Suzy does not.

Interestingly, Proposal 2 can also explain why PPC has some initial appeal. According to PPC, if an event $\vec{X} = \vec{x}$ prevents an event $\vec{E} = \vec{e}$, then $\vec{X} = \vec{x}$ must causally produce an alternative $\vec{E} = \vec{e'}$ to $\vec{E} = \vec{e}$. Proposal 2 basically admits that if $\vec{X} = \vec{x}$ prevents $\vec{C} = \vec{c}$ from producing $\vec{E}=\vec{e}$, then $\vec{X}=\vec{x}$ must indeed causally (co-)produce an alternative of some event that would otherwise have been causally produced by $\vec{C}=\vec{c}$. But this event does not have to be an alternative $\vec{E}=\vec{e'}$ of $\vec{E}=\vec{e}$. Instead, it can be any event, whose alternative would otherwise have been a part of the active causal path from $\vec{C}=\vec{c}$ to $\vec{E}=\vec{e}$. Still, there is a grain of truth in PPC: Whenever an event $\vec{X}=\vec{x}$ prevents an event $\vec{C}=\vec{c}$ from producing an effect $\vec{E}=\vec{e}$, there is some event $\vec{Y}=\vec{y}$ such that $\vec{X}=\vec{x}$ prevents $\vec{C}=\vec{c}$ from producing $\vec{Y}=\vec{y}$ by (co-)producing an alternative $\vec{Y}=\vec{y'}$ of $\vec{Y}=\vec{y}$. In many cases $\vec{Y}=\vec{y}$ is not identical to the effect of interest $\vec{E}=\vec{e}$. But there may very well be cases, in which $\vec{Y}=\vec{y}$ is identical to $\vec{E}=\vec{e}$. In these cases, PPC holds true. Imagine, for example, a gust of wind that prevents an archer from hitting bullseye with his arrow. The gust of wind co-produces (together with the archer's shot) an alternative impact of the arrow. Here, the preventer, namely the gust of wind, actually co-produces an alternative to the prevented event, which is the arrow hitting bullseye.[28]

This last point suggests that Proposal 2 still needs a slight modification, because the formulation of Pr2 in Proposal 2 does not allow that $\vec{Y}=\vec{y'}$ is identical to $\vec{E}=\vec{e}$. Adjusting this shortcoming leads to my final proposal for defining prevention:

Actual Prevention. *$\vec{X}=\vec{x}$ prevents $\vec{C}=\vec{c}$ from producing $\vec{E}=\vec{e}$ in the causal setting $(\mathcal{M},\vec{u},Prod)$ if and only if either (Pr1, Pr2, Pr3) or (Pr1, Pr2', Pr3) are satisfied:*

Pr1 $(\mathcal{M},\vec{u},Prod)\models\vec{X}=\vec{x}\wedge\vec{C}=\vec{c}\wedge\neg(\vec{C}=\vec{c}\textit{ produces }\vec{E}=\vec{e})$.

Pr2 *There is an event $\vec{Y}=\vec{y}$ such that there is a partition $(\vec{Z},\vec{W})$ of the endogenous Variables $\mathcal{V}$ in $\mathcal{M}$ with $\vec{X}\subseteq\vec{Z}$ and some setting $(\vec{x}',\vec{w})$ such that if $(\mathcal{M},\vec{u},Prod)\models Z=z^*$ for all $Z\in\vec{Z}$, then:*

(a) $(\mathcal{M},\vec{u},Prod)\models[\vec{X}\leftarrow\vec{x}',\vec{W}\leftarrow\vec{w}](\vec{C}=\vec{c}\textit{ produces }\vec{Y}=\vec{y'})\wedge(\vec{Y}=\vec{y'}\textit{ (co-)produces }\vec{E}=\vec{e})$.

(b) $(\mathcal{M},\vec{u},Prod)\models[\vec{X}\leftarrow\vec{x},\vec{W}'\leftarrow\vec{w},\vec{Z}'\leftarrow\vec{z^*}]\neg(\vec{C}=\vec{c}\textit{ produces }\vec{E}=\vec{e})\wedge(\vec{X}=\vec{x}$ *(co-)produces* $\vec{Y}=\vec{y})$ *for all* $\vec{W}'\subseteq\vec{W}$ *and for all* $\vec{Z}'\subseteq\vec{Z}$.

Pr2' *There is a partition $(\vec{Z},\vec{W})$ of the endogenous Variables $\mathcal{V}$ in $\mathcal{M}$ with $\vec{X}\subseteq\vec{Z}$ and some setting $(\vec{x}',\vec{w})$ such that if $(\mathcal{M},\vec{u},Prod)\models Z=z^*$ for all $Z\in\vec{Z}$, then:*

(a) $(\mathcal{M},\vec{u},Prod)\models[\vec{X}\leftarrow\vec{x}',\vec{W}\leftarrow\vec{w}](\vec{C}=\vec{c}\textit{ produces }\vec{E}=\vec{e})$.

(b) $(\mathcal{M},\vec{u},Prod)\models[\vec{X}\leftarrow\vec{x},\vec{W}'\leftarrow\vec{w},\vec{Z}'\leftarrow\vec{z^*}]\neg(\vec{C}=\vec{c}\textit{ produces }\vec{E}=\vec{e})\wedge(\vec{X}=\vec{x}$ *(co-)produces* $\vec{E}=\vec{e'})$ *for all* $\vec{W}'\subseteq\vec{W}$ *and for all* $\vec{Z}'\subseteq\vec{Z}$.

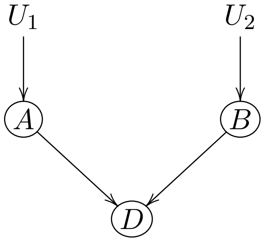
Pr3 *$\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}'=\vec{x}'$ (with $\vec{x}'$ being the restriction of $\vec{x}$ to the variables in $\vec{X}'$) satisfies conditions Pr1 and Pr2/Pr2'.*

Notice that the definition accords well with our expanded Power PC theory and the axioms about prevention therein: Axiom C10 resurfaces in condition Pr1 and axiom C11 is explicated in condition Pr2/Pr2'.

[28] This small example also illustrates why I use the term '(co-)production' in conditions Pr2(a) and Pr2(b), instead of 'production': $\vec{X}=\vec{x}$, for example, does not necessarily have to produce $\vec{Y}=\vec{y}$, it is sufficient if it co-produces $\vec{Y}=\vec{y}$ together with some other event.

6.5.3 A Problem with Abstract Causal Models

There is still a problem with our final definition of prevention. So far, I have simply presupposed that any causal model, that represents a case of prevention, is detailed enough to explicitly represent the respective variables of collision. But abstract representations of causal scenarios, that abstain from representing detailed causal processes, are quite common in the practice of causal modeling. It might therefore happen that a causal model represents a case of prevention without explicitly representing the respective variables of collision. Just consider an example about a poisoning, which turned out to be quite popular in the causal model literature: An assassin puts a lethal dose of poison into Pete's tea. Luckily, Pete's bodyguard is a very precautionary man. Out of a gut feeling he puts an antidote into the tea before Pete takes his first sip. The antidote completely neutralizes the poison and Pete survives. Figure 6.7 illustrates how the situation is typically modeled in the literature.[29] The bivalent variable A represents whether the assassin puts a lethal dose of poison into Pete's tea, the bivalent variable B represents whether Pete's bodyguard puts the antidote into Pete's tea, and the bivalent variable D represents whether Pete dies.



- $A := U_1$
- $B := U_2$
- $D := A \land \neg B$

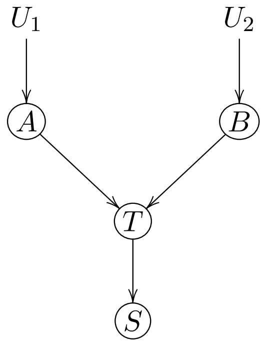
Figure 6.7: Causal model of the poisoning scenario.

Intuitively, the action of Pete's bodyguard ($B = 1$) clearly prevented the assassin ($A = 1$) from producing Pete's death ($D = 1$). But the bodyguard's action ($B = 1$) does not satisfy condition Pr2 in our definition of prevention, since there is no event $\vec{Y} = \vec{y}$ in the given causal model, that $B = 1$ (co-)produces and whose alternative $\vec{Y} = \vec{y}'$ would otherwise have been a part of an active causal path from the assassin's poisoning ($A = 1$) to Pete's death ($D = 1$). In short: there is no variable of collision for the bodyguard's prevention in the causal model.

But a more detailed description of the situation reveals that there is indeed a variable of collision. An antidote typically works like this: After the ingestion it produces antibodies in Pete's blood. If there are toxic molecules from the poison in Pete's bloodstream, then the antibodies bind to the toxic molecules and thereby change their form. Unbinded toxic molecules would quickly damage cells in Pete's body in such a drastic way that it would lead to Pete's death. But binded toxic molecules, due to their changed form, are unable to damage any cells in Pete's body. A more detailed causal model of the scenario would therefore look like the one

[29] See, for example, (Hiddleston, 2005), (Hitchcock, 2007), (Halpern and Hitchcock, 2015).

in Figure 6.8.



- $A := U_1$
- $B := U_2$

- $T = 0$ iff $A = 0$;
  $T = 1$ iff $A = 1 \wedge B = 0$;
  $T = 2$ iff $A = 1 \wedge B = 1$
- $D = 1$ iff $T = 1$

Figure 6.8: Expanded causal model of the poisoning scenario.

Besides the already known variables, the expanded model contains the additional variable T that can take on three different values. $T = 0$ represents that there are no toxins in Pete's bloodstream. $T = 1$ represents that there are non-binded toxins in Pete's bloodstream and $T = 2$ represents that there are binded toxins in Pete's bloodstream. The causal model in Figure 6.8 makes now explicit how $B = 1$ prevents $A = 1$ from producing $D = 1$ by explicitly incorporating the variable of collision T. If both $A = 1$ and $B = 1$ are the case, then $B = 1$ co-produces $T = 2$ together with $A = 1$. As a result, Pete survives. But if $B = 1$ is not the case, then $A = 1$ produces the alternative $T = 1$, which in turn causally produces Pete's death ($D = 1$). So, in the causal setting, in which $U_1 = U_2 = 1$, $B = 1$ now satisfies condition Pr2 in our definition of prevention, which now adequately recognizes that $B = 1$ prevents $A = 1$ from producing $D = 1$.

The example illustrates that the adequacy of our prevention definition depends on how detailed the causal model is, in which we apply the definition. There are at least two possible solutions to this problem, which I will call *the abstract model problem for prevention.* For the first solution, the following definition by Halpern (2016) proves to be helpful:[30]

Conservative Extension. *Given causal models $\mathcal{M}' = (\mathcal{U}', \mathcal{V}', \mathcal{R}', \mathcal{F}')$ and $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{R}, \mathcal{F})$, the causal setting $(\mathcal{M}', \vec{u}', Prod')$ is a conservative extension of the causal setting $(\mathcal{M}, \vec{u}, Prod)$ if and only if:*

(a) *$\mathcal{V} \subset \mathcal{V}'$.*

(b) *$Prod \subset Prod'$.*

(c) *For all variables $X \in \mathcal{V}$ and all settings $\vec{w}$ of the variables in $\vec{W} = V - \{X\}$: $(\mathcal{M}, \vec{u}) \models [\vec{W} \leftarrow \vec{w}]X = x$ iff $(\mathcal{M}', \vec{u}') \models [\vec{W} \leftarrow \vec{w}]X = x$*

[30] See (Halpern, 2016, p. 123). The definition given here differs from Halpern's original definition in so far that I consider causal models that are augmented by a set *Prod*, which entails information about production-relations. I have therefore added condition (b) to the definition.

A conservative extension of a causal setting $(\mathcal{M}, \vec{u}, \mathit{Prod})$ refines the representation of the represented causal scenario, but it conserves all the actual events, the counterfactual dependencies, and therefore the causal relationship assessments in $(\mathcal{M}, \vec{u}, \mathit{Prod})$.[31] With the concept of a conservative extension at hand, we could augment our definition of prevention in the following way:

Actual Prevention. *$\vec{X} = \vec{x}$ prevents $\vec{C} = \vec{c}$ from producing $\vec{E} = \vec{e}$ in the causal setting $(\mathcal{M}, \vec{u}, \mathit{Prod})$ if and only if there is an adequate conservative exstension $(\mathcal{M}', \vec{u}', \mathit{Prod}')$ of $(\mathcal{M}, \vec{u}, \mathit{Prod})$, such that according to our definition of prevention: $\vec{X} = \vec{x}$ prevents $\vec{C} = \vec{c}$ from producing $\vec{E} = \vec{e}$ in the causal setting $(\mathcal{M}', \vec{u}', \mathit{Prod}')$.*

The second solution to the abstract model problem for prevention is this: We introduce an adequacy or aptness constraint on causal models, according to which any causal model that represents a potential prevention-relation has to be detailed enough to explicitly include the respective variable(s) of collision. So, in the poisoning example from above, for example, only the second causal model that is illustrated in figure 6.8 would be acknowledged as an adequate representation of the given scenario.

This solution may at first sound somewhat radical, since we would brand many examples of causal models, that have been deployed in the causal model literature, as inadequate. This even includes two of my own previous examples: In the basketball scenario, as well as in the baseball scenario, I have deployed causal models that represent prevention relations without including the respective variables of collision. Still, the introduction of an aptness constraint on causal models is a legitimate move to solve the abstract model problem for prevention. It is without any question that modeling choices highly affect whether our formal definitions of causation will yield intuitive results for the modeled scenarios, and the 'art of modeling', as Halpern and Hitchcock (2010) call it, is clearly a domain that has not gotten sufficient attention in the causal model literature so far.[32] I am actually convinced that there are good arguments for introducing a general aptness constraint, according to which any causal model that represents a potential prevention relation should include the respective variable(s) of collision.[33] But it would go beyond the scope of this chapter to discuss these arguments in detail. In the following, I will nonetheless implement this aptness constraint for three simple reasons. First, it does no harm. The constraint does not lead us to inadequate, but only to more detailed causal models that represent prevention relations more thoroughly. Secondly, it solves the abstract model problem for prevention. And finally, as the next section will illustrate, models that satisfy this aptness constraint are especially helpful for differentiating the causal producer in a strong actual cause from those parts that are intuitively recognized as mere dependence-causes.

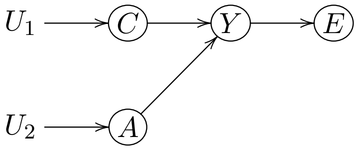
[31] That a conservative extension of a causal setting $(\mathcal{M}, \vec{u}, \mathit{Prod})$ conserves all the causal relationship assessments in $(\mathcal{M}, \vec{u}, \mathit{Prod})$ is only approximately true. Halpern (2016) shows that this depends on the respective definitions of causation. I will not go into the details of this, since I prefer the second solution to the abstract model problem for prevention anyhow.

[32] A few exceptions are (Halpern and Hitchcock, 2010), (Woodward, 2016), (Blanchard and Schaffer, 2017).

[33] Such a constraint provides, for example, a promising approach for dealing with several instances of the so-called *problem of isomorphism*. For discussions of the problem of isomorphism, see, for example, (Hitchcock, 2007) (Halpern and Hitchcock, 2015), (Menzies, 2017), or (Blanchard and Schaffer, 2017).

6.6 Sorting the Producer from the Chaff

With a definition of prevention at hand, we can now come back to the question that we have posed at the end of section 6.4: Can we formulate a condition that enables us to differentiate between those parts of a strong actual cause that are intuitively recognized as causal producers and those parts that are intuitively recognized as mere dependence-causes, like omissions of preventers? To illustrate the challenge that we are facing, consider the causal model in figure 6.9 which entails the bivalent variables C, A, Y and E:



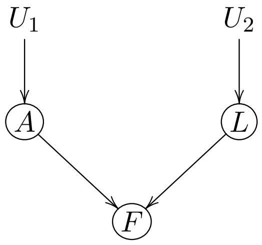
- $Y := C \wedge \neg A$
- $E := Y$

Figure 6.9: A causal model that may or may not entail a preventive cause.

Imagine a causal setting with $U_1 = 1$ and $U_2 = 0$. $C = 1 \wedge A = 0$ is then a strong actual cause of $E = 1$ in the given causal setting. Imagine further that 1 is a deviant state of E, which means that $E = 1$ is considered to be causally produced in the given setting. According to hypothesis 3, $C = 1 \wedge A = 0$ entails a causal producer of $E = 1$. But so far we cannot say what exactly the causal producer is. There are three possibilities: (1) $C = 1 \wedge A = 0$ is a causal producer of $E = 1$. (2) $C = 1$ is a causal producer of $E = 1$, while $A = 0$ is an omission of a ($C = 1$)-specific prevention of $E = 1$. (3) $A = 0$ is a causal producer of $E = 1$, while $C = 1$ is an omission of a ($A = 0$)-specific prevention of $E = 1$. The challenge that we are facing is whether there is any way to determine which of the three possibilities is correct, without presupposing the correct answer.

Here is what we can do: We have presupposed that $E = 1$ is in a deviant state, which means that $Y = 1$ must be a causal producer of $E = 1$. Therefore, $A = 0$ can only be a (co-)producer of $E = 1$, if it is a (co-)producer of $Y = 1$. And, according to our counterfactual theory of prevention, $A = 0$ can only be an omission of a ($C = 1$)-specific preventer of $E = 1$, if its alternative $A = 1$ would (co-)produce $Y = 0$, if $A = 1$ would be the case. The same holds for $C = 1$: $C = 1$ can only be a (co-)producer of $E = 1$, if it is a (co-)producer of $Y = 1$. And $C = 1$ can only be an omission of an ($A = 0$)-specific preventer of $E = 1$, if its alternative $C = 0$ would (co-)produce $Y = 0$, if it would be the case. This illustrates that, with our definition of prevention, the problem of differentiating between a (co-)producer and an omission of a preventer is reduced to the problem of differentiating between a (co-)producer and an omission of a (co-)producer. But this is a problem that we have already learned to deal with. As shown in section 6.4, we can differentiate between a (co-)producer and an omission of a (co-)producer by examining whether the effect in question is a default or a deviant state. So, this very same information will now also help us to differentiate between a (co-)producer and an omission of a preventer. Imagine, for example, that $Y = 1$ is a deviant state and $Y = 0$ a default state in the given causal setting. This means that $Y = 0$ is considered to be causally unproduced. According to our definition of prevention, neither $A = 1$ nor $C = 0$ could therefore be a preventer of $E = 1$, if they were the case. $A = 0$ and $C = 1$ are therefore no omissions of preventers of $E = 1$. Instead, both must be co-producers, which means that $C = 1 \wedge A = 0$ is a causal producer of $E = 1$.

Let us now take on some concrete examples. First, consider the conjunctive forest fire scenario from chapter 1 as illustrated in figure 6.10.



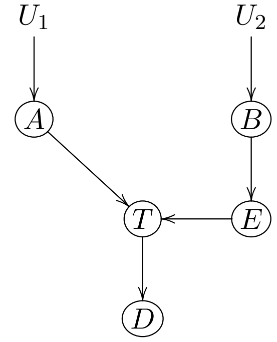
- $A := U_1$
- $L := U_2$
- $F := A \wedge L$

Figure 6.10: The conjunctive forest fire scenario.

Imagine a causal setting with $U_1 = U_2 = 1$. In that case $A = 1 \wedge L = 1$ is a strong actual cause of $F = 1$. 0 is considered to be the self-sustaining default value of F, while 1 is considered to be the deviant state of F. Given that the causal model is detailed enough, that it would entail the variable(s) of collision of any potential prevention-relation that it might represent, F is the only candidate for a variable of collision. Since 0 is considered to be the default value of F, there is no possible event in the causal model that, if it would be the case, would causally produce $F = 0$. From this it follows that neither $A = 1$ nor $L = 1$ can be an omission of a preventer of $F = 1$. Instead, both must be co-producers of $F = 1$, which means that $A = 1 \wedge L = 1$ is a causal producer of $F = 1$.

Next, consider the poisoning scenario as presented in figure 6.8. Imagine a causal setting with $U_1 = 1$ and $U_2 = 0$, which means that the assassin poisons Pete's tea ($A = 1$), the bodyguard does not put any antidote in it ($B = 0$), and Pete dies ($D = 1$). Applying our condition of strong actual causation, we get that $A = 1 \wedge B = 0$ is a strong actual cause of $D = 1$. 1 (dying) is considered to be the deviant state of D, while 0 (living) is considered to be the default state. $T = 1$ is therefore a causal producer of $D = 1$. Given that the causal model is detailed enough, that it would entail the variable(s) of collision of any potential prevention-relation that it might represent, T is the only candidate for a variable of collision. In the given situation, T has the value 1. According to our definition of prevention, $A = 1$ is therefore either a (co-)producer of $T = 1$ or an omission of a (co-)producer of an alternative of $T = 1$. The default state of T is 0 (not having any kind of toxins in the blood). 1 and 2 are both deviant states of T. $A = 0$ would yield $T = 0$, if it would be the case. But since $T = 0$ is a default state, $A = 0$ cannot be a causal (co-)producer of $T = 0$, which means that $A = 1$ is no omission of a (co-)producer of an alternative of $T = 1$. $A = 1$ is therefore a causal (co-)producer of $T = 1$ and, accordingly, a causal (co-)producer of $D = 1$. But what about $B = 0$? Here again, according to our definition of prevention, $B = 0$ is either a (co-)producer of $T = 1$ or an omission of a (co-)producer of an alternative of $T = 1$. $B = 1$ would yield $T = 2$, if it would be the case. But $T = 2$ is, just like $T = 1$, a deviant state. So, $B = 1$ could be a causal (co-)producer of $T = 2$, just like $B = 0$ could be a causal (co-)producer of $T = 1$. We have therefore no indication for deciding whether $B = 0$ is a (co-)producer of $T = 1$ or an omission of a (co-)producer of an alternative of $T = 1$. So, how can we explain our clear intuition that $B = 0$ is an omission of a (co-)producer of an alternative of $T = 1$ and therefore an omission of an $(A = 1)$-specific preventer of $D = 1$?

A more detailed representation of the scenario might be helpful. We intuitively know that $B = 1$ would not yield $T = 2$ directly, if it would be the case. When the Bodyguard puts an antidote in Pete's tea, which also entails a poison, this does not directly make Pete's blood contain bound toxins. First the molecules of the antidote must be ingested and they must enter the bloodstream. We can represent this mediating event by a bivalent variable E, representing whether the antibody molecules enter the bloodstream, and use the causal model shown in figure 6.11.



- $A := U_1$
- $B := U_2$
- $E := B$
- $T = 0$ iff $A = 0$;
  $T = 1$ iff $A = 1 \wedge E = 0$;
  $T = 2$ iff $A = 1 \wedge E = 1$
- $D = 1$ iff $T = 1$

Figure 6.11: An even more expanded causal model of the poisoning scenario.

Now, in the considered causal setting, we have $E = 0$, because of $B = 0$. And $B = 0$ can only influence T via E, which means that $B = 0$ can only be a causal (co-)producer of $T = 1$ if $B = 0$ is a causal producer of $E = 0$. But intuitively, 0 is the default state of E (antibody molecules do not enter Pete's bloodstream), while 1 is the deviant state of E (antibody molecules are entering Pete's bloodstream). $B = 0$ is therefore no causal producer of $E = 0$ and, accordingly, no causal (co-)producer of $T = 1$. $B = 1$, on the other hand, would be a causal producer of $E = 1$, if it would be the case, and therefore also a causal co-producer of $T = 2$. This makes $B = 0$ an omission of a co-producer of an alternative to $T = 1$ and therefore, according to our definition of prevention, an omission of an $(A = 1)$-specific preventer of $D = 1$.

It is time to take stock. We have already argued in section 6.4 that, whether an event $Y = y$ is considered to have a causal producer in a given causal setting $(\mathcal{M}, \vec{u})$, can be decided by differentiating between default and deviant values of Y in $(\mathcal{M}, \vec{u})$. Also, whenever $Y = y$ does have a causal producer in $(\mathcal{M}, \vec{u})$, then any strong actual cause of $Y = y$ in $(\mathcal{M}, \vec{u})$ entails exactly one causal producer of $Y = y$. But besides a causal producer, a strong actual cause of $Y = y$ may also entail mere-dependence causes of $Y = y$, like omissions of preventers. In the present section we have seen that, given our counterfactual account of prevention and given that the causal model under consideration is sufficiently detailed to entail the variable(s) of collision of any potential prevention-relation, that the represented scenario might entail, the distinction between default and deviant states can be sufficient for differentiating between those parts of a strong actual cause that are intuitively recognized as causal producers and those parts that are intuitively recognized as mere dependence-causes.

6.7 Summary

In the present chapter, I have illustrated that the concepts of causal production and causal prevention, both concepts of token causation that are crucial ingredients of our expanded Power PC theory, differ intensionally and extensionally from interventionist concepts of token causation, like actual, sufficient, or strong actual causation. This is in accord with Hall's (2004) position, that there are two reasonable, but decisively distinct conceptions of causation: dependence- and production-causation. We noticed that, while every causal producer of a given effect is also a dependence-cause of the effect, not every dependence-cause of the effect is also a causal producer. In chapters 1 and 2, I have already presented precise criteria for identifying certain dependence-causes in the framework of causal models. But in the present chapter we realized, that we do not yet have clear-cut criteria that enable the differentiation between causal producers and mere dependence-causes. I have argued that the most prominent acounts of production-causation in the literature so far, namely Dowe's (2000) CQ theory and Glennan's (2017) mechanistic account, are unable to provide us with such criteria.

Building on Dowe's counterfactual theory of prevention, I have put forward a new account of prevention that reduces the concept of prevention to counterfactual claims about causal production in the framework of causal models. Based on this reduction, I have then proposed an account, according to which the interventonist concept of strong actual causation and the differentiation between default and deviant states in a causal model are often sufficient to differentiate causal producers from mere dependence-causes. Whether these tools will be sufficient for reliably identifying the causal producers of a given effect in any kind of causal scenario, remains to be shown. But I hope to have illustrated that there is a promising approach to analyze the concept of causal production without relying on any kind of locality thesis. My proposed account of causal production is promising, because it adequately captures several intuitions that previous accounts of production-causation are unable to explain. For example, as several examples by Schaffer (2004) illustrate, our intuitions of causal production often seem to be dependent on how a causal scenario is framed or modeled. Analyzing the concept of causal production in terms of the model-relative concept of strong actual causation and the model-relative distinction between default and deviant states can explain this phenomenon. Furthermore, we often have clear intuitions about causal production in fictitious examples, in which information about any mechanisms or physical processes between cause and effect are completely missing. If, for example, in some magical world a wizard wields his magic wand and makes an object appear out of thin air, then we have the clear intuition that the wizard just causally produced the appearance of the object, even though we have no information about the underlying physical processes or mechanisms. Or if God creates light by simply saying "Let there be light", we have the clear intuition that God causally produced the light, even though we do not have any information about the underlying physical processes or mechanisms. What we do have, though, are intuitions about the default and the deviant states in the given fictitious scenarios. We also have intuitions about the counterfactual dependencies. According to my newly proposed approach, these information can be sufficient for establishing intuitions about causal production.

One might worry that analyzing the concept of causal production by presupposing a distinction between default and deviant states might end up being circular. If the concept of a default state is defined as a state that is causally unproduced, then for knowing whether something is a default state, we would have to know, if some event causally produces it. But, as accounts by Maudlin (2004) or Hitchcock (2007) suggest, there are promising paths to identify default states in a given system without relying on a previous identification of causal productions. Still, for a comprehensive account of causal production that is based on the distinction between default and deviant states, a more detailed analysis of this distinction would be instructive. But as already pointed out, it has not been the goal of this chapter to provide a complete or comprehensive analysis of causal production. Instead, the main objective of this chapter has been to obtain new insights about how the concepts of causal production, causal prevention, strong actual causation and the distinction between default and deviant states all interrelate. In the next chapter, I will put these newly gained insights to use for one of the main goals of this dissertation: A formal explication of causal explanation.

Chapter 7

Extensive Causal Explanations

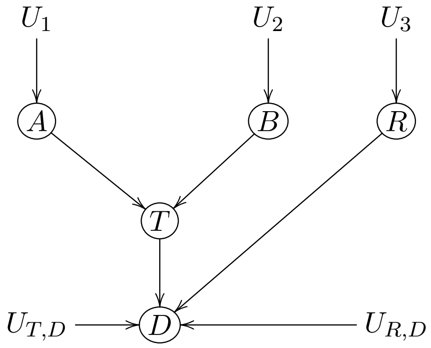
7.1 Introduction

The definitions of causal explanation, as developed in chapters 1 and 3, are solely based on an interventionist concept of dependence-causation, namely strong actual causation. The concepts of causal production and causal prevention do, so far, not play any role in our explications of explanations. But, as philosophers like Dowe (2000) and Hall (2004) argue and psychologists like Cheng (1997) and Walsh and Sloman (2011) corrobarate, the concepts of causal production and causal prevention play a key role in our causal understanding of the world. It should therefore not be surprising that the concepts of causal production and prevention, as well as the differentiation between causal producers and mere dependence-causes, also have an influence on our intuitions about causal explanations. In the present chapter, I aim to explore how this influence looks like.

In the process of this, we will see that, what I have described in chapters 1 and 3 as complete and explicitly complete causal explanations, sometimes still lack explanatory relevant information. Incorporating this additional information will lead us to, what I will call, *extensive causal explanations*. These extensive causal explanations exhibit an internal structure that our considerations about causal explanations in chapters 1 and 3 have not yet revealed. This internal structure is essentially governed by the differentiation between generative and preventive causes. In accordance with an observation that has already been made by Paul Humphreys (1989), I will argue that, in general, any extensive causal explanation of an explanandum ϕ has the following form: 'ϕ, *because* $\vec{X} = \vec{x}$, *despite* $\vec{Z} = \vec{z}$'. Of what kind of causal factors the *because*- and the *despite*-part are composed, crucially depends on whether the explanandum ϕ is understood to be a deviant or a default state.

7.2 Extensive Causal Explanations of Deviant States

Consider the following variation of the poisoning scenario from section 6.5: An assassin poisons Pete's tea ($A = 1$) with a poison that has an intrinsic causal power of 0.9 to produce death, once it has entered the bloodstream and as long as the blood does not contain any antidote molecules. But Pete's bodyguard puts an antidote into Pete's tea ($B = 1$) whenever he has a gut feeling that he should do so. The antidote is perfectly reliable and would therefore prevent the poison from killing Pete with a preventive power of 1. But since Pete has many enemies, it is also possible that there is another assassin, acting independently of the first, who contemplates on shooting Pete with a rifle ($R = 1$). Statistics on the shooting abilities of the second assassin tell us that $R = 1$ has a causal power of 0.7 on Pete's death ($D = 1$). The probabilistic SEM in figure 7.1 represents the described scenario.



- $A := U_1$
- $B := U_2$
- $R := U_3$

- $T = 0$ iff $A = 0$;
  $T = 1$ iff $A = 1 \wedge B = 0$;
  $T = 2$ iff $A = 1 \wedge B = 1$
- $D = 1$ iff $(T = 1 \wedge U_{T,D} = 1) \vee (R = 1 \wedge U_{R,D} = 1)$

Figure 7.1: Probabilistic variant of the poisoning scenario.

Imagine that our preliminary beliefs about the actions of the bodyguard and the two assassins in a given token situation is best described by the following probability assignments: $\mathcal{P}_{pre}(U_1 = 1) = \mathcal{P}_{pre}(U_2 = 1) = \mathcal{P}_{pre}(U_3 = 1) = 0.5$. According to our knowledge about the intrinsic generative causal powers of $A = 1$ and $R = 1$ on $D = 1$, we have: $\mathcal{P}_{pre}(U_{T,D} = 1) = 0.9$, and $\mathcal{P}_{pre}(U_{R,D} = 1) = 0.7$. Imagine further, that we come to learn that Pete actually dies ($D = 1$). With $\vec{u} = (u_1, u_2, u_3, u_{T,D}, u_{R,D})$ we therefore have the following contexts $\vec{u}_i$ with a probability $\mathcal{P}(\vec{u}_i) = \mathcal{P}_{pre}(\vec{u}_i | D = 1) > 0$:[1]

- $\vec{u}_1 = (1, 1, 1, 1, 1)$
- $\vec{u}_2 = (1, 1, 1, 0, 1)$
- $\vec{u}_3 = (1, 0, 1, 0, 1)$
- $\vec{u}_4 = (1, 0, 1, 1, 1)$
- $\vec{u}_5 = (1, 0, 1, 1, 0)$
- $\vec{u}_6 = (0, 1, 1, 1, 1)$
- $\vec{u}_7 = (0, 1, 1, 0, 1)$
- $\vec{u}_8 = (0, 0, 1, 0, 1)$
- $\vec{u}_9 = (0, 0, 1, 1, 1)$
- $\vec{u}_{10} = (1, 0, 0, 1, 1)$
- $\vec{u}_{11} = (1, 0, 0, 1, 0)$

Now, what are the potential explanations of Pete's death in the given probabilistic causal model? As a quick reminder, here is how we have defined the concept of an explicitly complete explanation in chapter 3:

Explicitly Complete Explanation. $\vec{X} = \vec{x}$ *is an explicitly complete explanation of* ϕ *relative to* $\mathcal{K}^{\mathcal{P}}$ *in* $\mathcal{M}$ *if and only if* $\mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} \phi) > 0$.

[1] In this and in the following examples, the exact probability values will be irrelevant, which is why I will not list them in detail.

Applying this definition in our example gives us the following explicitly complete explanations of the deviant state $D = 1$:[2]

(1) $A = 1 \wedge B = 0$, because it is a strong actual cause of $D = 1$ in $\vec{u}_4$, $\vec{u}_5$, $\vec{u}_{10}$, $\vec{u}_{11}$.

(2) $R = 1$, because it is a strong actual cause of $D = 1$ in $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_3$, $\vec{u}_4$, $\vec{u}_6$, $\vec{u}_7$, $\vec{u}_8$, $\vec{u}_9$.

According to hypothesis 3 from chapter 6, every explicitly complete explanation of a deviant state ϕ entails exactly one event that has a non-zero probability of being a causal producer of ϕ.[3] I will call such an event a *potential causal producer* of ϕ. The potential causal producer of $D = 1$ in $A = 1 \wedge B = 0$ is $A = 1$ and the potential causal producer of $D = 1$ in $R = 1$ is, non surprisingly, $R = 1$. As is also evident from the last chapter, an explicitly complete causal explanation of a deviant state ϕ can also contain more than just a potential causal producer of ϕ. It can additionally contain events that have a certain probability of being mere dependence-causes of ϕ. But no explicitly complete causal explanation of a deviant state ϕ can ever do without a potential causal producer of ϕ. It is therefore fair to say that a potential causal producer forms the fundamental pillar of any explicitly complete causal explanation of a deviant state.

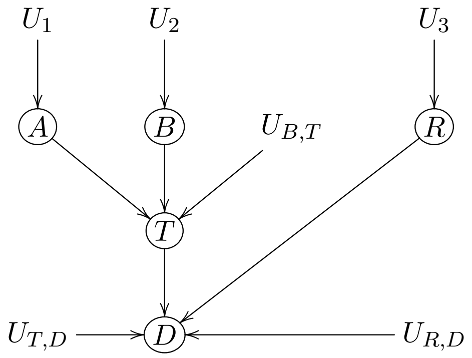
Now, besides a potential causal producer $\vec{C} = \vec{c}$, an explicitly complete causal explanation of a deviant state ϕ additionally contains any factor on which the causal production of ϕ by $\vec{C} = \vec{c}$ might de facto depend. In our example, $B = 1$ would prevent the causal production of $D = 1$ by $A = 1$, if it would be the case. The causal production of $D = 1$ by $A = 1$ is therefore de facto dependent on $B = 0$. The information, that $B = 0$ is the case, is therefore clearly explanatory relevant for any explanation that cites $A = 1$ as a potential causal producer of $D = 1$. In the given example, the definition of an explicitly complete explanation therefore does an intuitively satisfying job in providing us with explanations of the explanandum that entail all explanatory relevant information. But we only need to make a slight variation of the example to see that this is not always the case.

Imagine the same causal scenario as before, only this time $B = 1$ is not perfectly reliable as an $(A = 1)$-specific preventer of $D = 1$. Instead, it only has an $(A = 1)$-specific preventive power of 0.8 on $D = 1$, which means that, given $A = 1$, $B = 1$ causally co-produces $T = 2$ with a probability of 0.8. This gives us the probabilistic causal model shown in figure 7.2.

We again assume a preliminary distribution with $\mathcal{P}_{pre}(U_1 = 1) = \mathcal{P}_{pre}(U_2 = 1) = \mathcal{P}_{pre}(U_3 = 1) = 0.5$. According to our causal power assumptions, we additionally have: $\mathcal{P}_{pre}(U_{T,D} = 1) = 0.9$, $\mathcal{P}_{pre}(U_{B,T} = 1) = 0.8$, and $\mathcal{P}_{pre}(U_{R,D} = 1) = 0.7$. Imagine again, that we come to learn that Pete actually dies ($D = 1$). With $\vec{u} = (u_1, u_2, u_3, u_{B,T}, u_{T,D}, u_{R,D})$ we have the following contexts $\vec{u}_i$ with a probability $\mathcal{P}(\vec{u}_i) = \mathcal{P}_{pre}(\vec{u}_i | D = 1) > 0$:

[2] Here and in the following examples I will not explicitly state the proofs for claims about relations of strong actual causation. The claims can easily be verified by checking the fulfillment of the conditions provided in the definition of strong actual causation as presented in chapter 1.

[3] Notice that, according to hypothesis 3, every strong actual cause of a deviant state ϕ entails exactly one causal producer of ϕ. Now, an explicitly complete explanation of ϕ is an event with a non-zero probability of being a strong actual cause of ϕ. This is why, according to hypothesis 3, every explicitly complete explanation of ϕ entails exactly one event that has a non-zero probability of being a causal producer of ϕ.



- $A := U_1$
- $B := U_2$
- $R := U_3$

- $T = 0$ iff $A = 0$;
  $T = 1$ iff $A = 1 \wedge \neg(B = 1 \wedge U_{B,T} = 1)$;
  $T = 2$ iff $A = 1 \wedge (B = 1 \wedge U_{B,T} = 1)$
- $D = 1$ iff $(T = 1 \wedge U_{T,D} = 1) \vee (R = 1 \wedge U_{R,D} = 1)$

Figure 7.2: The probabilistic poisoning scenario with an unreliable antidote.

- $\vec{u}_1 = (1, 1, 1, 1, 1, 1)$
- $\vec{u}_2 = (1, 1, 1, 0, 1, 1)$
- $\vec{u}_3 = (1, 1, 1, 0, 1, 0)$
- $\vec{u}_4 = (1, 1, 1, 0, 0, 1)$
- $\vec{u}_5 = (1, 1, 1, 1, 0, 1)$
- $\vec{u}_6 = (1, 0, 1, 1, 1, 1)$
- $\vec{u}_7 = (1, 0, 1, 0, 1, 1)$
- $\vec{u}_8 = (1, 0, 1, 0, 1, 0)$
- $\vec{u}_9 = (1, 0, 1, 0, 0, 1)$
- $\vec{u}_{10} = (1, 0, 1, 1, 0, 1)$
- $\vec{u}_{11} = (1, 0, 1, 1, 1, 0)$
- $\vec{u}_{12} = (0, 1, 1, 1, 1, 1)$
- $\vec{u}_{13} = (0, 1, 1, 0, 1, 1)$
- $\vec{u}_{14} = (0, 1, 1, 0, 0, 1)$
- $\vec{u}_{15} = (0, 1, 1, 1, 0, 1)$
- $\vec{u}_{16} = (0, 0, 1, 1, 1, 1)$
- $\vec{u}_{17} = (0, 0, 1, 0, 1, 1)$
- $\vec{u}_{18} = (0, 0, 1, 0, 0, 1)$
- $\vec{u}_{19} = (0, 0, 1, 1, 0, 1)$
- $\vec{u}_{20} = (1, 1, 0, 0, 1, 1)$
- $\vec{u}_{21} = (1, 1, 0, 0, 1, 0)$
- $\vec{u}_{22} = (1, 0, 0, 1, 1, 1)$
- $\vec{u}_{23} = (1, 0, 0, 0, 1, 1)$
- $\vec{u}_{24} = (1, 0, 0, 0, 1, 0)$
- $\vec{u}_{25} = (1, 0, 0, 1, 1, 0)$

We get the following explicitly complete explanations of the deviant state $D = 1$:

(1) $A = 1 \wedge B = 0$, because it is a strong actual cause of $D = 1$ in $\vec{u}_6$, $\vec{u}_{11}$, $\vec{u}_{22}$, $\vec{u}_{25}$.

(2) $A = 1$, because it is a strong actual cause of $D = 1$ in $\vec{u}_2$, $\vec{u}_3$, $\vec{u}_7$, $\vec{u}_8$, $\vec{u}_{20}$, $\vec{u}_{21}$, $\vec{u}_{23}$, $\vec{u}_{24}$.

(3) $R = 1$, because it is a strong actual cause of $D = 1$ in $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_4$, $\vec{u}_5$, $\vec{u}_6$, $\vec{u}_7$, $\vec{u}_9$, $\vec{u}_{10}$, $\vec{u}_{12}$, $\vec{u}_{13}$, $\vec{u}_{14}$, $\vec{u}_{15}$, $\vec{u}_{16}$, $\vec{u}_{17}$, $\vec{u}_{18}$, $\vec{u}_{19}$.

We can see that we now get two explicitly complete explanations of $D = 1$ with the same causal producer, namely $A = 1$. One of these explanations contains, besides $A = 1$, an information that, if true, would increase the probability that $A = 1$ causally produces $D = 1$, namely the fact that the only event, which is able to prevent $A = 1$ from producing $D = 1$, is absent. Intuitively,

$A = 1 \wedge B = 0$ is therefore a potential explanation of $D = 1$ that does not lack any explanatory relevant information.

But what about the second explicitly complete explanation, the one that only contains the causal producer $A = 1$? Here, explanatory relevant information is missing. Whether $B = 0$ or $B = 1$ is the case, is explanatory relevant to any explanation that cites $A = 1$ as a potential causal producer of $D = 1$, because it has a significant impact on the probability that $A = 1$ successfully produces $D = 1$. Imagine that Pete's wife learns that Pete died ($D = 1$) and she wants to know why. Her epistemic state is represented by the probabilistic SEM in figure 7.2. When we tell her that an assassin poisened Pete's tea ($A = 1$), she will understandably be discontent with that explanation. She would probably feel pressed to ask a clarificatory question: "But what about Pete's bodyguard?". She knows that Pete's bodyguard occasionally puts an antidote in Pete's tea which significantly lowers a poisons capacity to kill. Knowing the value of B, that is, knowing whether the bodyguard has put the antidote into Pete's tea or not, is therefore highly relevant for assessing how powerful the cause $A = 1$ really is with respect to Pete's death. So, as long as Pete's wife does not know the value of B, she is unable to assess how good, or powerful $A = 1$ is as a causal explanation of Pete's death.

Admittedly, we have not yet discussed the concept of explanatory power. I will do so in the next chapter. For now, I simply presuppose, what I consider to be a highly intuitive claim anyhow, namely that the explanatory power of a causal explanation of a deviant state ϕ is determined by the effective causal influence of the cited potential causal producer of ϕ.[4] I will call any causal explanation of an explanandum ϕ that contains all the information that is needed to assess its own explanatory causal power, an *extensive causal explanation*. Since, only the value of B is needed to determine the effective causal influence of $A = 1$ on $D = 1$, and therefore the explanatory power of an explanation that cites $A = 1$ as a cause of $D = 1$, the explicitly complete explanation $A = 1 \wedge B = 0$ is an extensive causal explanation of $D = 1$, while the explicitly complete explanation $A = 1$ is no extensive causal explanation of $D = 1$. Instead, $A = 1 \wedge B = 1$ amounts to an extensive causal explanation of $D = 1$, even though it does not have a non-zero probability of being a strong actual cause of $D = 1$ and is therefore no explicitly complete explanation of $D = 1$.[5]

Still, we have to acknowledge that there is a significant difference between $A = 1 \wedge B = 0$ and $A = 1 \wedge B = 1$ as extensive causal explanations of $D = 1$. In both cases, we obtain all the information that is needed to assess the effective causal influence of $A = 1$ on $D = 1$ by learning the actual value of B. But in the first case, the information that $B = 0$ holds, increases the probability that $A = 1$ causally produces $D = 1$. It therefore increases the power of $A = 1$ as a causal explanation of $D = 1$. The information that $B = 1$ holds, on the other hand, decreases the probability that $A = 1$ causally produces $D = 1$. It therefore decreases the power of $A = 1$ as a causal explanation of $D = 1$. This suggests that extensive causal explanations of a given explanandum ϕ do, in general, have an internal structure that is expressed by the following

[4] Although I do consider this claim to be highly intuitive, I do not simply take it for granted. In chapter 8, I will provide arguments for its adequacy.

[5] $A = 1 \wedge B = 1$ is also no part of a strong actual cause of $D = 1$ in any causal setting of the given model, which means that it is also not recognized as a potential, actual or partial explanation of $D = 1$ by our definitions in chapter 3.

schema:[6]

$$\phi \textit{ because } \vec{X} = \vec{x}, \textit{ despite } \vec{Z} = \vec{z} \tag{7.1}$$

If ϕ is a deviant state, then the *because*-part $\vec{X} = \vec{x}$ of the explanation, contains one potential causal producer $\vec{C} = \vec{c}$ of ϕ and events that, if true, would increase the effective causal influence of $\vec{C} = \vec{c}$ on ϕ, like omissions of events that might prevent $\vec{C} = \vec{c}$ from producing ϕ. The *despite*-part $\vec{Z} = \vec{z}$, on the other hand, contains events that, if true, would decrease the effective causal influence of $\vec{C} = \vec{c}$ on ϕ, like events that are able to prevent $\vec{C} = \vec{c}$ from producing ϕ. Notice, though, that the *despite*-part of an extensive causal explanation of a deviant state ϕ may very well be empty. This is the case with the extensive causal explanation $A = 1 \wedge B = 0$ of $D = 1$ in the previous example, since, if $A = 1 \wedge B = 0$ is the case, then there is no factor left in the causal model that might decrease the effective causal influence of $A = 1$ on $D = 1$. If we express the explanation in the form of the general schema, we therefore get: '$D = 1$ *because* $A = 1 \wedge B = 0$'. The extensive causal explanation $A = 1 \wedge B = 1$ of $D = 1$, on the other hand, does have a non-empty *despite*-part: '$D = 1$ *because* $A = 1$, *despite* $B = 1$'.

7.3 An Algorithm for Generating Extensive Causal Explanations

The previous example has shown that the concept of an extensive causal explanation does not coincide with the definition of an explicitly complete explanation or with any of the explanation-concepts that we have defined in chapter 3. We do therefore not yet have a formal explication of our newly introduced concept of an extensive causal explanation. In this section, I aim to change that by providing an algorithm that generates only and all the extensive causal explanations of a given explanandum ϕ in a probabilistic SEM $(\mathcal{M}, \mathcal{P})$ in the form of the general schema 'ϕ *because* $\vec{X} = \vec{x}$, *despite* $\vec{Z} = \vec{z}$'.

To do so, I first need to introduce some new terminology. Let us say that $\vec{X} = \vec{x}$ is a *potential strong actual cause* of ϕ relative to a causal model $\mathcal{M}$ if and only if there is some logically possible context $\vec{u}$ for $\mathcal{M}$ such that $\vec{X} = \vec{x}$ is a strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$. We can then say:

Maximal Potential Strong Actual Cause. *$\vec{X} = \vec{x}$ is a 'maximal potential strong actual cause' of ϕ relative to $\mathcal{M}$ if and only if $\vec{X} = \vec{x}$ is a potential strong actual cause of ϕ relative to $\mathcal{M}$ and there is no $\vec{Y} = \vec{y}$ with $\vec{X} = \vec{x}$ being a strictly smaller part of $\vec{Y} = \vec{y}$ such that $\vec{Y} = \vec{y}$ is a potential strong actual cause of ϕ relative to $\mathcal{M}$.*

We can now define:

Extensive Causal Explanation. *Let $(\mathcal{M}, \mathcal{P})$ be a probabilistic causal model and let ϕ be an event in $\mathcal{M}$. Then for any $\vec{X} = \vec{x}$ with $\mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} \phi) > 0$:*

(a) *If $\vec{X} = \vec{x}$ is a maximal potential strong actual cause of ϕ relative to $\mathcal{M}$, then $\vec{X} = \vec{x}$ is an extensive causal explanation of ϕ of the following form: 'ϕ because $\vec{X} = \vec{x}$'.*

[6] As pointed out above, Paul Humphreys (1989) already argues for this.

(b) *If $\vec{X} = \vec{x}$ is not a maximal potential strong actual cause of ϕ relative to $\mathcal{M}$, then let $(\vec{X} = \vec{x} \wedge \vec{Z}^i = \vec{z}^i)_{i \in \{1,\ldots,m\}}$ be the family of all maximal potential strong actual causes of ϕ relative to $\mathcal{M}$, of which $\vec{X} = \vec{x}$ is a strictly smaller part. If $\mathcal{P}(\vec{X} = \vec{x} \wedge \vec{Z}^1 \neq \vec{z}^1 \wedge \ldots \wedge \vec{Z}^m \neq \vec{z}^m) > 0$,[7] then: $\vec{X} = \vec{x} \wedge \vec{Z}^1 \neq \vec{z}^1 \wedge \ldots \wedge \vec{Z}^m \neq \vec{z}^m$ is an extensive causal explanation of ϕ of the following form: 'ϕ because $\vec{X} = \vec{x}$, despite $\vec{Z}^1 \neq \vec{z}^1 \wedge \ldots \wedge \vec{Z}^m \neq \vec{z}^m$'.*

(c) *These are the only extensive causal explanations of ϕ in $(\mathcal{M}, \mathcal{P})$.*

The basic idea underlying this formal explication is this: We know that every potential strong actual cause of a deviant state ϕ contains exactly one causal producer $\vec{C} = \vec{c}$ of ϕ. So, whenever we have a potential strong actual cause $\vec{X} = \vec{x}$ of ϕ that is a strictly smaller part of a maximal potential strong actual cause $\vec{X} = \vec{x} \wedge \vec{Z} = \vec{z}$ of ϕ, then we know that $\vec{X} = \vec{x}$ contains a causal producer $\vec{C} = \vec{c}$ of ϕ, while $\vec{Z} = \vec{z}$ does not contain any causal producer of ϕ. Instead, it only contains events that, if true, would increase the probability that $\vec{C} = \vec{c}$ causally produces ϕ, like omissions of events that are able to prevent $\vec{C} = \vec{c}$ from producing ϕ. Now, since rules (a) and (b) both demand that the *because*-part of an extensive causal explanation of ϕ always consists of a potential strong actual cause of *phi*, both rules ensure that any extensive causal explanation of a deviant state ϕ contains a causal producer $\vec{C} = \vec{c}$ of ϕ in the *because*-part. The same demand also ensures that in addition to a causal producer $\vec{C} = \vec{c}$ of ϕ, the *because*-part of an extensive causal explanation of ϕ may also contain events that, if true, would increase the probability that $\vec{C} = \vec{c}$ causally produces ϕ. Any maximal potential strong actual cause of ϕ that cites $\vec{C} = \vec{c}$ as a causal producer of ϕ contains the omission (or a preventer) of any event that might prevent $\vec{C} = \vec{c}$ from producing ϕ. This is why, according to rule (a), any maximal potential strong actual cause $\vec{X} = \vec{x}$ of ϕ amounts to an extensive causal explanation of ϕ that has $\vec{X} = \vec{x}$ in the *because*-part and an empty *despite*-part. But if a potential strong actual cause of ϕ that cites $\vec{C} = \vec{c}$ as a causal producer of ϕ is non-maximal, then it lacks information about the values of variables, that might turn out to be potential $(\vec{C} = \vec{c})$-specific preventers of ϕ. Rule (b) basically adds this missing information, by claiming that these potential $(\vec{C} = \vec{c})$-specific preventers of ϕ are present. Rule (b) thereby produces all the extensive causal explanations of ϕ that are no maximal potential strong actual causes of ϕ.

All these considerations are rather abstract. So, to provide a more concrete illustration of how our formal explication of extensive causal explanations works, let us consider some examples. Let us start with the poisoning scenario in figue 7.2. We have already pointed out that we have three events $\vec{X} = \vec{x}$ with $\mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} D = 1) > 0$:

- $A = 1 \wedge B = 0$,
- $A = 1$
- $R = 1$

First, $A = 1 \wedge B = 0$ is a maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$. So, according to rule (a), we get the following extensive causal explanation of $D = 1$ in $(\mathcal{M}, \mathcal{P})$:

[7] $\vec{Z}^i \neq \vec{z}^i$ is short for $\neg(Z^i_1 = z^i_1) \wedge \ldots \wedge \neg(Z^i_n = z^i_n)$.

(1) $D = 1$ *because* $A = 1 \wedge B = 0$.

The despite-part of (1) is empty, while the *because*-part contains a potential causal producer of $D = 1$ and the omission of an event that, if true, would be able to prevent this potential causal producer from producing $D = 1$: Pete died because an assassin has put poison into his tea and Pete's bodyguard did not put any anitode into his tea. Next, $R = 1$ is also a maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$. So, according to rule (a), we get the following extensive causal explanation of $D = 1$ in $(\mathcal{M}, \mathcal{P})$:

(2) $D = 1$ *because* $R = 1$.

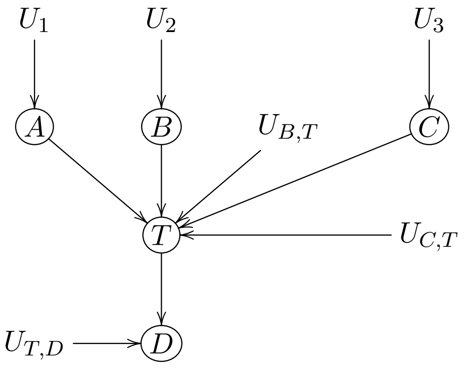
The *despite*-part of (2) is empty, while the *because*-part contains a potential causal producer of $D = 1$: Pete died because an assassin has shot him. Finally, $A = 1$ is no maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$. $A = 1 \wedge B = 0$ is the only maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$, of which $A = 1$ is a strictly smaller part. Since $\mathcal{P}(A = 1 \wedge B = 1) = \mathcal{P}(\{\vec{u}_1, \vec{u}_2, \vec{u}_3, \vec{u}_4, \vec{u}_5, \vec{u}_{20}, \vec{u}_{21}\}) > 0$, we get, according to rule (b), the following extensive causal explanation of $D = 1$ in $(\mathcal{M}, \mathcal{P})$:

(3) $D = 1$ *because* $A = 1$, *despite* $B = 1$.

The *because*-part of (3) contains a potential causal producer of $D = 1$, while the *despite*-part contains an event that is able to prevent this potential causal producer from producing $D = 1$: Pete died because an assassin has put poison into his tea, despite the fact that his bodyguard has put an antidote into his tea. (3) is just the extensive causal explanation of Pete's death, that the concept of an explicitly complete explanation was unable to capture. (3) is not very powerful as an explanation of Pete's death, since $A = 1 \wedge B = 1$, if true, do not make Pete's death highly probable. But (3) is nonetheless extensive, in the sense that it provides us with all the information that we need to assess its explanatory power. According to rule (c), there are no further extensive causal explanations of $D = 1$ in $(\mathcal{M}, \mathcal{P})$ besides (1), (2), and (3).

Let us consider yet another variation of the poisoning scenario. This time, there is no danger of a second assassin who might shoot Pete. Instead, we have two potential preventers of Pete's poisoning. Imagine that Pete's cook, who prepares Pete's tea is also aware that his boss is at constant danger of being poisoned. This is why he too has developed the habit to put an antidote into Pete's tea, whenever his gut tells him to do so. But the cook's antidote works differently than the bodyguard's antidote. So, whether the cook's antidote successfully prevents the poison from killing Pete is completely independent of whether the bodyguard's antidote, if administered, does so. Also, the cook's antidote is a bit less effective, since it only has an $(A = 1)$-specific preventive power of 0.5 on $D = 1$. Let the bivalent variable C represent whether Pete's cook puts his antidote into Pete's tea, with the value 1 representing that he does and the value 0 representing that he does not do so. The probabilistic SEM in figure 7.3 represents the scenario.

We again assume a preliminary distribution with $\mathcal{P}_{pre}(U_1 = 1) = \mathcal{P}_{pre}(U_2 = 1) = \mathcal{P}_{pre}(U_3 = 1) = 0.5$. According to our causal power assumptions, we additionally have: $\mathcal{P}_{pre}(U_{T,D} = 1) = 0.9$, $\mathcal{P}_{pre}(U_{B,T} = 1) = 0.8$, and $\mathcal{P}_{pre}(U_{C,T} = 1) = 0.5$. Imagine again, that we come to learn that



- $A := U_1$
- $B := U_2$
- $R := U_3$

- $T = 0$ iff $A = 0$;
  $T = 1$ iff $A = 1 \wedge \neg(B = 1 \wedge U_{B,T} = 1) \wedge \neg(C = 1 \wedge U_{C,T} = 1)$;
  $T = 2$ iff $A = 1 \wedge [(B = 1 \wedge U_{B,T} = 1) \vee (C = 1 \wedge U_{C,T} = 1)]$
- $D = 1$ iff $T = 1 \wedge U_{T,D} = 1$

Figure 7.3: The probabilistic poisoning scenario with two unreliable antidotes.

Pete actually dies ($D = 1$). With $\vec{u} = (u_1, u_2, u_3, u_{T,D}, u_{B,T}, u_{C,T})$ we then have the following contexts $\vec{u}_i$ with a probability $\mathcal{P}(\vec{u}_i) = \mathcal{P}_{pre}(\vec{u}_i | D = 1) > 0$:

- $\vec{u}_1 = (1, 1, 1, 1, 0, 0)$
- $\vec{u}_2 = (1, 1, 0, 1, 0, 1)$
- $\vec{u}_3 = (1, 1, 0, 1, 0, 0)$
- $\vec{u}_4 = (1, 0, 1, 1, 1, 0)$
- $\vec{u}_5 = (1, 0, 1, 1, 0, 0)$
- $\vec{u}_6 = (1, 0, 0, 1, 0, 0)$
- $\vec{u}_7 = (1, 0, 0, 1, 1, 0)$
- $\vec{u}_8 = (1, 0, 0, 1, 0, 1)$
- $\vec{u}_9 = (1, 0, 0, 1, 1, 1)$

The following events have a non-zero probability of being a strong actual cause of $D = 1$ in the given probabilistic SEM $(\mathcal{M}, \mathcal{P})$:

- $A = 1$, because it is a strong actual cause of $D = 1$ in $\vec{u}_1$, $\vec{u}_3$, $\vec{u}_5$, $\vec{u}_6$.
- $A = 1 \wedge B = 0$, because it is a strong actual cause of $D = 1$ in $\vec{u}_4$, $\vec{u}_7$.
- $A = 1 \wedge C = 0$, because it is a strong actual cause of $D = 1$ in $\vec{u}_2$, $\vec{u}_8$.
- $A = 1 \wedge B = 0 \wedge C = 0$, because it is a strong actual cause of $D = 1$ in $\vec{u}_9$.

Since $A = 1 \wedge B = 0 \wedge C = 0$ is a maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$, we get the following extensive causal explanation of $D = 1$ in $(\mathcal{M}, \mathcal{P})$ according to rule (a):

(1) $D = 1$ *because* $A = 1 \wedge B = 0 \wedge C = 0$.

The *despite*-part of (1) is empty, while the *because*-part contains a potential causal producer of $D = 1$ and the omission of the two events that, if true, would be able to prevent this potential causal producer from producing $D = 1$: Pete died because an assassin has put poison into his tea and both his bodyguard and his cook did not put any antidotes into his tea.

$A = 1 \wedge B = 0$ is no maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$. $A = 1 \wedge B = 0 \wedge C = 0$ is the only maximal potential strong actual cause of $D = 1$ in $\mathcal{M}$, of which $A = 1 \wedge B = 0$ is a strictly smaller part. Since $\mathcal{P}(A = 1 \wedge B = 0 \wedge C = 1) = \mathcal{P}(\{\vec{u}_4, \vec{u}_5\}) > 0$, we get, according to rule (b), the following extensive causal explanation of $D = 1$ in $(\mathcal{M}, \mathcal{P})$:

(2) $D = 1$ *because* $A = 1 \wedge B = 0$, *despite* $C = 1$.

The *because*-part of (2) contains a potential causal producer of $D = 1$ and the omission of an event that, if true, would be able to prevent this potential causal producer from producing $D = 1$. The *despite*-part of (2) contains an event that is able to prevent the potential causal producer from producing $D = 1$: Pete died because an assassin has put poison into his tea and his bodyguard did not put an antidote into his tea, despite the fact that his cook has put an antidote into his tea.

$A = 1 \wedge C = 0$ is no maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$. $A = 1 \wedge B = 0 \wedge C = 0$ is the only maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$, of which $A = 1 \wedge C = 0$ is a strictly smaller part. Since $\mathcal{P}(A = 1 \wedge C = 0 \wedge B = 1) = \mathcal{P}(\{\vec{u}_2, \vec{u}_3\}) > 0$, rule (b) gives us the following extensive causal explanation of $D = 1$ in $(\mathcal{M}, \mathcal{P})$:

(3) $D = 1$ *because* $A = 1 \wedge C = 0$, *despite* $B = 1$.

The *because*-part of (3) contains a potential causal producer of $D = 1$ and the omission of an event that, if true, would be able to prevent this potential causal producer from producing $D = 1$. The *despite*-part of (2) contains an event that is able to prevent the potential causal producer from producing $D = 1$: Pete died because an assassin has put poison into his tea and his cook did not put an antidote into his tea, despite the fact that his bodyguard has put an antidote into his tea.

Finally, $A = 1$ is no maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$. $A = 1 \wedge B = 0 \wedge C = 0$ is the only maximal potential strong actual cause of $D = 1$ relative to $\mathcal{M}$, of which $A = 1$ is a strictly smaller part. Since $\mathcal{P}(A = 1 \wedge C = 1 \wedge B = 1) = \mathcal{P}(\vec{u}_1) > 0$, rule (b) gives us the following extensive causal explanation of $D = 1$ in $(\mathcal{M}, \mathcal{P})$:

(4) $D = 1$ *because* $A = 1$, *despite* $B = 1 \wedge C = 1$.

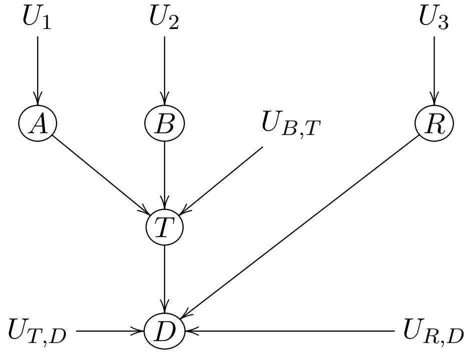
The *because*-part of (4) contains a potential causal producer of $D = 1$, while the *despite*-part contains the two events that are able to prevent the potential causal producer from producing $D = 1$: Pete died because an assassin has put poison into his tea, despite the fact that both his bodyguard and his cook have put their antidotes into Pete's tea.

According to rule (c), there are no further extensive causal explanations of $D = 1$ in $(\mathcal{M}, \mathcal{P})$ besides (1), (2), (3) and (4).

7.4 Extensive Causal Explanations of Default States

It is now time to focus on the explanation of default states. Since a default state is understood to be causally unproduced in the confines of the given causal model, it is clear that the *because*-part of an extensive causal explanation of a default state ϕ does not contain any potential causal producer of ϕ. Instead, it may contain omissions of events that have the intrinsic causal power to causally produce an alternative to ϕ or events that have the intrinsic power to prevent other events from causally producing an alternative to ϕ. The *despite*-part, on the other hand, may contain potential causal producers of alternatives to ϕ and omissions of events that have the intrinsic power to prevent other events from causally producing an alternative to ϕ.

As an illustration, consider again the probabilistic poisoning scenario with an unreliable antidote, shown again in figure 7.4, in which Pete faces the danger of two potential assassins, while his bodyguard might try to protect him from the first assassin with an antidote:



- $A := U_1$
- $B := U_2$
- $R := U_3$

- $T = 0$ iff $A = 0$;
$T = 1$ iff $A = 1 \wedge \neg(B = 1 \wedge U_{B,T} = 1)$;
$T = 2$ iff $A = 1 \wedge (B = 1 \wedge U_{B,T} = 1)$
- $D = 1$ iff $(T = 1 \wedge U_{T,D} = 1) \vee (R = 1 \wedge U_{R,D} = 1)$

Figure 7.4: The probabilistic poisoning scenario with an unreliable antidote.

We again assume a preliminary distribution with $\mathcal{P}_{pre}(U_1 = 1) = \mathcal{P}_{pre}(U_2 = 1) = \mathcal{P}_{pre}(U_3 = 1) = 0.5$ and our causal power assumptions give us: $\mathcal{P}_{pre}(U_{T,D} = 1) = 0.9$, $\mathcal{P}_{pre}(U_{B,T} = 1) = 0.8$, and $\mathcal{P}_{pre}(U_{R,D} = 1) = 0.7$. But this time, we assume that we come to learn that Pete does not die ($D = 0$). With $\vec{u} = (u_1, u_2, u_3, u_{B,T}, u_{T,D}, u_{R,D})$ we then have the following contexts $\vec{u}_i$ with a probability $\mathcal{P}(\vec{u}_i) = \mathcal{P}_{pre}(\vec{u}_i | D = 0) > 0$:

- $\vec{u}_1 = (0, 1, 0, 1, 1, 1)$
- $\vec{u}_2 = (0, 1, 0, 0, 1, 1)$
- $\vec{u}_3 = (0, 1, 0, 0, 1, 0)$
- $\vec{u}_4 = (0, 1, 0, 0, 0, 1)$
- $\vec{u}_5 = (0, 1, 0, 0, 0, 0)$
- $\vec{u}_6 = (0, 1, 0, 1, 1, 0)$
- $\vec{u}_7 = (0, 1, 0, 1, 0, 1)$
- $\vec{u}_8 = (0, 1, 0, 1, 0, 0)$
- $\vec{u}_9 = (0, 0, 0, 1, 1, 1)$
- $\vec{u}_{10} = (0, 0, 0, 0, 1, 1)$
- $\vec{u}_{11} = (0, 0, 0, 0, 1, 0)$
- $\vec{u}_{12} = (0, 0, 0, 0, 0, 1)$
- $\vec{u}_{13} = (0, 0, 0, 0, 0, 0)$
- $\vec{u}_{14} = (0, 0, 0, 1, 1, 0)$
- $\vec{u}_{15} = (0, 0, 0, 1, 0, 1)$
- $\vec{u}_{16} = (0, 0, 0, 1, 0, 0)$
- $\vec{u}_{17} = (1, 1, 0, 1, 1, 1)$
- $\vec{u}_{18} = (1, 1, 0, 0, 0, 1)$
- $\vec{u}_{19} = (1, 1, 0, 0, 0, 0)$
- $\vec{u}_{20} = (1, 1, 0, 1, 1, 0)$
- $\vec{u}_{21} = (1, 1, 0, 1, 0, 1)$
- $\vec{u}_{22} = (1, 1, 0, 1, 0, 0)$
- $\vec{u}_{23} = (1, 0, 0, 0, 0, 1)$
- $\vec{u}_{24} = (1, 0, 0, 0, 0, 0)$
- $\vec{u}_{25} = (1, 0, 0, 1, 0, 1)$
- $\vec{u}_{26} = (1, 0, 0, 1, 0, 0)$
- $\vec{u}_{27} = (0, 1, 1, 0, 1, 0)$
- $\vec{u}_{28} = (0, 1, 1, 0, 0, 0)$
- $\vec{u}_{29} = (0, 1, 1, 1, 1, 0)$
- $\vec{u}_{30} = (0, 1, 1, 1, 0, 0)$
- $\vec{u}_{31} = (0, 0, 1, 0, 1, 0)$
- $\vec{u}_{32} = (0, 0, 1, 0, 0, 0)$
- $\vec{u}_{33} = (0, 0, 1, 1, 1, 0)$
- $\vec{u}_{34} = (0, 0, 1, 1, 0, 0)$
- $\vec{u}_{35} = (1, 1, 1, 0, 0, 0)$
- $\vec{u}_{36} = (1, 1, 1, 1, 1, 0)$
- $\vec{u}_{37} = (1, 1, 1, 1, 0, 0)$
- $\vec{u}_{38} = (1, 0, 1, 0, 0, 0)$
- $\vec{u}_{39} = (1, 0, 1, 1, 0, 0)$

Again, we first identify the events that have a non-zero probability of being a strong actual cause of the default state $D = 0$:

- $A = 0$, because it is a strong actual cause of $D = 0$ in $\vec{u}_3$, $\vec{u}_6$, $\vec{u}_{11}$, $\vec{u}_{14}$, $\vec{u}_{27}$, $\vec{u}_{29}$, $\vec{u}_{31}$, $\vec{u}_{33}$.
- $R = 0$, because it is a strong actual cause of $D = 0$ in $\vec{u}_4$, $\vec{u}_7$, $\vec{u}_{12}$, $\vec{u}_{15}$, $\vec{u}_{18}$, $\vec{u}_{21}$, $\vec{u}_{23}$, $\vec{u}_{25}$
- $B = 1$, because it is a strong actual cause of $D = 0$ in $\vec{u}_{20}$, $\vec{u}_{36}$.
- $R = 0 \wedge B = 1$, because it is a strong actual cause of $D = 0$ in $\vec{u}_{17}$.
- $A = 0 \wedge R = 0$, because it is a strong actual cause of $D = 0$ in $\vec{u}_1$, $\vec{u}_2$, $\vec{u}_9$, $\vec{u}_{10}$.

Since $A = 0 \wedge R = 0$ and $R = 0 \wedge B = 1$ are both maximal potential strong actual causes of $D = 0$ relative to $\mathcal{M}$, we get the following extensive causal explanations of $D = 0$ in $(\mathcal{M}, \mathcal{P})$ according to rule (a):

(1) $D = 0$ *because* $A = 0 \wedge R = 0$.

(2) $D = 0$ *because* $B = 1 \wedge R = 0$.

(1) contains two omissions of potential producers of the alternative $D = 1$ of $D = 0$ in its *because*-part: Pete did not die because the first assassin did not put any poison into Pete's tea and the second assassin did not shoot Pete. (2) contains one omission of a potential producer of $D = 1$, namely $R = 0$, and one event, namely $B = 1$, that is able to prevent another potential producer, namely $A = 1$, from producing $D = 1$: Pete did not die because his bodyguard has

put an antidote into his tea and the second assassin did not shoot on him. The *despite*-part is empty in both cases.

Next, $B = 1$ is no maximal potential strong actual cause of $D = 0$ relative to $\mathcal{M}$. $R = 0 \wedge B = 1$ is the only maximal potential strong actual cause of $D = 0$ relative to $\mathcal{M}$, of which $B = 1$ is a strictly smaller part. Since $\mathcal{P}(B = 1 \wedge R = 1) > 0$, rule (b) gives us the following extensive causal explanation of $D = 0$ in $(\mathcal{M}, \mathcal{P})$:

(3) $D = 0$ *because* $B = 1$, *despite* $R = 1$.

The *because*-part of (3) again contains an event that is able to prevent a potential causal producer, namely $A = 1$, from producing $D = 1$, while the *despite*-part of (3) contains a potential causal producer of $D = 1$: Pete did not die because his bodyguard has put an antidote into his tea and despite the fact that the second assassin has shot on Pete.

Next, $A = 0$ is no maximal potential strong actual cause of $D = 0$ relative to $\mathcal{M}$. $A = 0 \wedge R = 0$ is the only maximal potential strong actual cause of $D = 0$ relative to $\mathcal{M}$, of which $A = 0$ is a strictly smaller part. Since $\mathcal{P}(A = 0 \wedge R = 1) > 0$, rule (b) gives us the following extensive causal explanation of $D = 0$ in $(\mathcal{M}, \mathcal{P})$:

(4) $D = 0$ *because* $A = 0$, *despite* $R = 1$.

The *because*-part of (4) contains an omission of a potential causal producer, namely $A = 1$, of $D = 1$, while the *despite*-part contains a potential causal producer of $D = 1$: Pete did not die because the first assassin did not put any poison into his tea and despite the fact that the second assassin has shot on Pete.

Finally, $R = 0$ is no maximal potential strong actual cause of $D = 0$ relative to $\mathcal{M}$. $A = 0 \wedge R = 0$ and $R = 0 \wedge B = 1$ are the only maximal potential strong actual causes of $D = 0$ relative to $\mathcal{M}$, of which $A = 0$ is a strictly smaller part. Since $\mathcal{P}(R = 0 \wedge A = 1 \wedge B = 0) > 0$, rule (b) gives us the following extensive causal explanation of $D = 0$ in $(\mathcal{M}, \mathcal{P})$:

(5) $D = 0$ *because* $R = 0$, *despite* $A = 1$ *and* $B = 0$.

The *because*-part of (5) contains an omission of a potential causal producer, namely $R = 1$, of $D = 1$, while the *despite*-part contains a potential causal producer of $D = 1$ and the omission of an event that has an intrinsic power to prevent this potential producer from producing $D = 1$: Pete did not die because the second assassin did not shoot on him and despite the fact that the first assassin has put poison into his tea and his bodyguard did not put any antidote into his tea.

Do we now have covered all extensive causal explanations of $D = 0$ in the given probabilistic SEM $(\mathcal{M}, \mathcal{P})$? Not necessarily. Notice that $\mathcal{P}$ ascribes non-zero probabilities to the contexts $\vec{u}_{38}$ and $\vec{u}_{39}$, in which $A = 1$, $B = 0$, and $R = 1$ are the case, while the explanandum $D = 0$ still holds. In these contexts, there seems to be no explanation of $D = 0$, since all the facts speak against $D = 0$. Nonetheless, if we allow that the cause-candidate $\vec{X} = \vec{x}$ in the definition of strong actual causation may also be an empty conjunction, then the empty conjunction is acknowledged as a strong actual cause of $D = 0$ in $(\mathcal{M}, \vec{u}_{38})$ and $(\mathcal{M}, \vec{u}_{39})$: We can vary the values of all variables in $\mathcal{M}$ aside from D by interventions, but the value of D will not change as a consequence of these interventions in $(\mathcal{M}, \vec{u}_{38})$ and $(\mathcal{M}, \vec{u}_{39})$. The empty conjunction is therefore a potential strong actual cause of $D = 0$ relative to $\mathcal{M}$. And since the empty conjunction is a strictly smaller part of every maximal potential strong actual cause of $D = 0$, rule (b) gives us the following extensive causal explanation of $D = 0$ in $(\mathcal{M}, \mathcal{P})$:

(6) $D = 0$, *despite* $A = 1 \wedge B = 0 \wedge R = 1$.

(6) might not be what we typically like to call an explanation, since it only cites events that make the explanandum less probable. But all these events are nonetheless explanatory relevant. (6) makes explicit that $D = 0$ might happen, even though everything speaks against it and that, in such a situation, Pete has nobody to thank, but pure luck itself.

Now, according to rule (c), there are no further extensive causal explanations of $D = 0$ in $(\mathcal{M}, \mathcal{P})$ besides (1), (2), (3), (4), (5), and (6).

Before I conclude this section, I want to address a potential worry. One might want to object that (2) and (3) are no extensive causal explanations of $D = 0$ after all, because they lack information about the value of A and it seems that the value of A influences the power of an explanation of $D = 0$, which cites $B = 1$ as a cause of $D = 0$. Take, for example, explanation (2), which is '$D = 0$ *because* $B = 1 \wedge R = 0$'. If $A = 1$ is the case, then the explanatory power of (2) clearly depends on the $(A = 1)$-specific preventive power of $B = 1$ on $D = 1$. The higher the power of $B = 1$ to prevent $A = 1$ from producing $D = 1$, the better $D = 0$ is explained by $B = 1 \wedge R = 0$. But if $A = 0$ is the case, then the $(A = 1)$-specific preventive power of $B = 1$ on $D = 1$ is completely irrelevant for the explanatory power of (2) on $D = 0$, because there is nothing that $B = 1$ is able to prevent in the first place. But this point already hints at a potential rebuttel of the objection: The value of A is not really relevant for the explanatory power of an explanation of $D = 0$, which cites $B = 1$ as a cause of $D = 0$. Instead, the value of A is only relevant for whether an explanation candidate, which cites $B = 1$ as a cause of $D = 0$, is an explanation of $D = 0$ in the first place. $B = 1$ can only be (part of) a strong actual cause of $D = 0$ in a given causal setting, if $A = 1$ also holds in this setting. So, whenever we claim that $B = 1$ is (part of) a strong actual cause of $D = 0$, then this claim already entails the claim that $A = 1$ holds. Accordingly, explaining $D = 0$ by, for example, claiming that $B = 1 \wedge R = 0$ are causes of $D = 0$ entails the claim that $A = 1$ holds. The extensive causal explanations (2) and (3) do therefore not really lack any explanatory relevant information, since they both imply the claim that $A = 1$ holds.

Still, one might have the intuition that an extensive causal explanation should include such highly relevant information not only implicitly, but explicitly. In that case, one could amend the explication of extensive causal explanations from section 7.3 by adding a rule that basically says: Whenever the *because*-part of an extensive causal explanation of a default state ϕ contains an event that has an intrinsic power to prevent another event $\vec{C} = \vec{c}$ from producing an alternative to ϕ, then $\vec{C} = \vec{c}$ must be incorporated into the *despite*-part of the explanation. According to this rule, the extensive causal explanation (2), for example, would be supplemented to: '$D = 0$ *because* $B = 1 \wedge R = 0$, *despite* $A = 1$'. I consider this amendment to be legitimate. Still, I am not quite sure whether I share the intuition that motivates it. Since the amendment would also make the algorithm, that produces extensive causal explanations for a given explanandum, somewhat more complex, I will continue to adhere to the algorithm as presented in section 7.3. But this decision is mostly based on pragmatic considerations. Opting for the amendment seems to be an equally appropriate way to go.

7.5 Summary

In the present chapter, I have argued that an explicitly complete causal explanation of an explanandum ϕ, as defined in chapter 3, does not always contain all the information that is needed to evaluate its own explanatory power on ϕ. I have therefore put forward a new concept of causal explanation that is able to fullfill this demand, namely the concept of an extensive causal explanation. I have argued that, in general, any extensive causal explanation of an explanandum ϕ has an internal structure that is expressed by the the following schema: 'ϕ, *because* $\vec{X} = \vec{x}$, *despite* $\vec{Z} = \vec{z}$'. Additionally, I have explored how the differentiation between the *because*- and the *despite*-part of an extensive causal explanation of an explanandum ϕ is affected by the differentiation between generative and preventive causes of ϕ, depending on whether ϕ is understood to be a default or a deviant state.

Chapter 8

Explanatory Power

8.1 Introduction

By now, I have introduced several distinct, but nonetheless related, definitions of causal explanation. The multiplicity of concepts nicely reflects our multifaceted intuitions about causal explanations. But there is still a certain complexity in our intuitions about causal explanations, that I have largely neglected so far. According to the definitions, that I have put forward in chapters 1, 3, and 7, a given event either is a potential/actual/correct/explicitly complete/extensive causal explanation of a given explanandum or it is not. A more nuanced grading is, so far, not taken into account. But intuitively, we have a strong tendency of grading causal explanations on a fine, potentially even continuous, scale, once we have identified them. There is only one concept of causal explanation, for which we have already taken this tendency into consideration, namely the concept of partial explanation. In chapters 1 and 3, I have put forward the definition of a partial explanation with a corresponding measure of explanatory completeness. Some (partial) explanations seem to be more complete or closer to completeness than other partial explanations, and the proposed measure is able to capture these intuitions. But it is not just partial explanations that we tend to rate on a certain scale or spectrum. Even extensive causal explanations, all of which are as complete as causal explanations can get, are clearly scattered on a wide spectrum of explanatory excellence or goodness: Some extensive causal explanations just seem to do a better job in explaining a given explanandum than other extensive causal explanations. They are simply more explanatory.

In the present chapter, I aim to explicate how explanatory goodness can be measured. Or at least a certain dimension of explanatory goodness. Because a closer look reveals that there are actually several distinct ways, in which one explanation can be better or more explanatory than another.

8.2 Dimensions of Explanatory Goodness

8.2.1 Unification and Coherence

The variety of accounts that aim to describe what exactly explanations and their defining characteristics are has brought with it a variety of different proposals of what exactly explanatory goodness is. Take, for example, unificationist accounts of explanation, like the ones developed by (Friedman, 1974), (Kitcher, 1981), (Kitcher, 1989), (Bartelborth, 2002), or (Schurz and Lambert, 1994). According to unificationist accounts, an hypothesis H explains a given explanandum E, if E is deducible from H and H unifies a previously incoherent set of theories, beliefs or phenomena. This suggests a straightforward criterion of explanatory goodness: One explanation candidate can be better than another by yielding a bigger increase of unification or coherence.[1]

The account of causal explanation that I have put forward in this dissertation is clearly not built on a unificationist conception of explanation. The ability to unify a set of theories, beliefs or phenomena is, according to my approach, no defining feature of a causal explanation.[2] But it may still be the case that the ability to unify a certain set of beliefs is one way in which a causal explanation can be better than its alternatives. Imagine that we have two potential causal explanations, H_1 and H_2, of a given explanandum E. While H_2 can only explain E, H_1 is able to explain E as well as several other events that have been observed and that are still in need of causal explanations. In that case, H_1 may indeed have an explanatory advantage over H_2. So, even though the ability to unify is, according to my account of causal explanations, no defining feature of a causal explanation, the ability to unify a certain set of beliefs may still be a dimension of explanatory goodness.

8.2.2 Explanatory Depth

Hitchcock and Woodward (2003) put forward another conception of explanatory goodness. Here again, Hitchcock and Woodward's interpretation of explanatory goodness is closely connected with their account of causal explanation, which is spelled out in (Woodward and Hitchcock, 2003) and more extensively in (Woodward, 2003). According to this account, a causal explanation of a given explanandum ϕ consists of a description of certain token events I_1, ..., I_n, which serve as initital conditions, and some causal generalizations, according to which the initial conditions I_1, ..., I_n may causally yield the explanandum ϕ.[3] Now, Hitchcock and Woodward (2003) argue that a causal explanation H_1 is, ceteris paribus, better than another causal explanation H_2, if the causal generalizations in H_1 are more stable or invariant than the causal generalizations in H_2. Hitchcock and Woodward call this dimension of explanatory goodness *explanatory depth*. As Hitchcock and Woodward (2003) point out, there are actually several ways, in which a causal generalization can be more invariant than another, which basically gives us several subdimensions of explanatory depth.[4]

[1] See, for example, (Thagard, 1989) and (Glass, 2007) for how to measure and compare the ability of explanations to increase coherence.

[2] This does not exclude the possibility, though, that there are indeed certain non-causal explanations, for which unification is indeed a defining feature.

[3] Just like Halpern and Pearl (2005b), Woodward (2003) uses the framework of causal models to explicate his account of causal explanation. In his account, the initial conditions I_1, ..., I_n are therefore represented as events in a causal model, while the causal generalizations are represented as structural equations. Woodward's account of causal explanation therefore has some crucial similarities to Halpern and Pearl's (2005b) account and to the account developed in this dissertation. But there are also some crucial differences. For example, Woodward's conception of causal explanation is not agent-dependent, that is, explanations are not defined as being relative to some agent's epistemic state.

[4] For example, a causal generalization G_1 may be more invariant than another causal generalization G_2 under interventions on the variables that are explicitly mentioned in G_1 and G_2. But at the same time G_1 may be less invariant than G_2 under changes of background conditions that are not explicitly mentioned in G_1 and G_2.

As Hitchcock and Woodward (2003) point out, the explanatory depth of a causal explanation is often closely connected with its degree of specificity or abstractness. A very specific causal generalization, which describes a causal relationship on a very detailed level, is, for example, typically highly invariant under changes of background conditions. This is why, subsequently the label of *explanatory depth* has often been used to describe the degree of specificity or abstractness of a causal explanation.[5] Both antagonistic values have, at least to a certain degree, some intuitive appeal as explanatory virtues: While very specific explanations tend to contain more invariant causal relationships, abstract explanations may ignore irrelevant details and they typically encompass more singular instances than more specific explanations. More abstract explanations therefore seem to have an advantage when it comes to simplification and unification.

8.2.3 Explanatory Power

Another dimension of explanatory goodness is associated with an understanding of explanations that especially Hempel's *DN*- and *IS*-account have advocated: A successful explanation should make the explanandum expectable. Based on this conviction, Schupbach and Sprenger (2011) have developed a measure of explanatory goodness, that quantifies an explanation's "ability to decrease the degree to which we find the explanandum surprising" (Schupbach and Sprenger, 2011, p. 108). This dimension of explanatory goodness has been discussed in the philosophical literature under the heading of *explanatory power*.

Since the concepts of causal explanation, that I have developed in this dissertation, are always relative to a given epistemic state and since the belief in the presence of a cause tends to increase the belief in the presence of its potential effects, the concept of explanatory power is a well-suited dimension of explanatory goodness to evaluate the explanations that are encompassed by my definitions. In the following sections, I will discuss how explanatory power can be quantified. I will not deal with the virtues of unification or coherence and the virtue of explanatory depth. This is not because I think that these dimensions of explanatory goodness are irrelevant for our concepts of causal explanation. They may very well be intuitively adequate and normatively useful dimensions of the goodness of a causal explanation.[6] The reason that I focus on explanatory power is instead, that the account of causal explanation, which I have developed in this dissertation, proofs to be especially well-equipped to advance our understanding of explanatory power and to explicate how it can be measured.

8.3 Three Bayesian Measures of Explanatory Power

Different explanations of a given explanandum ϕ can vary widely in their ability to reduce an agent's surprise about ϕ. Just take the poisoning scenario from figure 7.2 as an example. Imagine that we come to learn that Pete has died, but we do not yet know why. Coming to learn, that an assassin has put a poison into Pete's tea while Pete's bodyguard has not put any antidote into the tea, strongly reduces the surprise about Pete's death. Coming to learn, that

[5] See, for example, (Weslake, 2013), (Strevens, 2008), (Kinney, 2019a).

[6] This is why, in chapters 9 and 10, I will deal with a question that is closely connected to the concept of explanatory depth, namely the question of whether there is some ideal level of abstraction for a causal explanation.

an assassin has put a poison into Pete's tea while Pete's bodyguard has put an antidote into the tea, also reduces the surprise about Pete's death, but it does so significantly less. The second extensive causal explanation is therefore much less powerful than the first. But how exactly can we quantify an explanations ability to reduce an agent's surprise about a given explanandum? In the present section, I will discuss three different proposals, that have been developed in the context of Bayesian probability theory and that have become predominant in the philosophical literature:[7] The measure ε_{SS}, that has been put forward by Schupbach and Sprenger (2011), the measure ε_{GMG}, which has been proposed by Good (1960) and McGrew (2003), and the measure ε_{CT}, proposed by Crupi and Tentori (2012).

All three measures have been characterized and motivated in terms of certain adequacy constraints and corresponding representation theorems. By now, the literature has produced a number of different representation theorems, especially for ε_{SS} and ε_{CT}, that are able to characterize the measures in terms of different adequacy constraints.[8] I will base the following discussion on the representation theorems by Sprenger and Hartmann (2019), since the employed adequacy constraints allow a straightforward comparison of all three measures.

ε_{SS}, ε_{GMG}, and ε_{CT} all aim to quantify the ability of an explanation H to reduce an agent's surprise about a given explanandum E. For doing so, all three measures are based on the assumption that an agent's epistemic state can be represented by a probability distribution P over a set of propositions. Given such a probability distribution P, an agent's degree of surprise about an explanandum E can be represented as the inverse of its probability $P(E)$. The lower the probability that an agent α assigns to E, the more surprising is the occurrence of E to α. Accordingly, an explanation H is able to lower α's surprise about E, if learning H increases the probability that α assigns to E, that is, if $P(E|H) > P(E)$. This idea can be summarized by the following adequacy constraint that all three measures comply with:[9]

Generalized Difference Making - Explanatory Power (GDM-EP). *There is a real-valued, continuous function $f : [0,1]^2 \mapsto \mathbb{R}$ such that for a putative explanation H and a putative explanandum E, we can write the explanatory power $\varepsilon(H,E)$ of H on E as:*

$$\varepsilon(H,E) = f(P(E|H), P(E)) \tag{8.1}$$

where f is non-decreasing in the first argument and non-increasing in the second argument.

After having described where all three measures agree, let us now see where they come apart. For this, consider the following adequacy constraint for a measure ε of explanatory power:[10]

Confirmatory Value (CV). $\varepsilon(H,E_1) > \varepsilon(H,E_2)$ *if and only if* $P(H|E_1) > P(H|E_2)$

CV says that an explanation H explains an explanandum E_1 better than an alternative explanandum E_2 if and only if the posterior probability of H after learning E_1 is higher than the posterior probability of H after learning E_2. Notice that a similar condition, typically under the

[7] This is why, in the following, I will also refer to the three measures as *the Bayesian measures of explanatory power.*

[8] See, for example, (Schupbach and Sprenger, 2011), (Schupbach, 2011b), (Crupi and Tentori, 2012), (Cohen, 2015), (Cohen, 2016), (Sprenger and Hartmann, 2019).

[9] See (Sprenger and Hartmann, 2019, p. 193).

[10] See (Sprenger and Hartmann, 2019, p. 195).

name of *Final Probability Incrementality*, is a common constraint on any confirmation-measure c: $c(H, E_1) > c(H, E_2)$ if and only if $P(H|E_1) > P(H|E_2)$.[11] Taking this into account, CV basically says that H explains an explanandum E_1 better than an alternative explanandum E_2 if and only if E_1 confirms H better than E_2 does. CV therefore puts explanatory power into a close relationship with confirmation. Now, Sprenger and Hartmann (2019, cf. p. 195) prove the following representation theorem for ε_{GMG}:

Representation Theorem for ε_{GMG}. *All measures of explanatory power $\varepsilon(H, E)$ that satisfy GDM-EP and CV are ordinally equivalent to $\varepsilon_{GMG}(H, E)$:*

$$\varepsilon_{GMG}(H, E) = log\frac{P(E|H)}{P(E)} \tag{8.2}$$

A simple application of Bayes' theorem gives us:

$$\frac{P(E|H)}{P(E)} = \frac{P(H|E)}{P(H)} \tag{8.3}$$

This reveals that ε_{GMG} is actually identical to a famous confirmation measure, namely the *Log-Ratio Measure*:[12]

$$c(H, E) = log\frac{P(H|E)}{P(H)} \tag{8.4}$$

Accepting both measures would therefore mean that the explanatory power of H on E is just the same as the degree of confirmation that learning E confers on H.[13] Arguing against a strict identification of explanatory power with degree of confirmation, Crupi (2012) shows that any measure of explanatory power that satisfies CV cannot also satisfy the following adequacy constraint:

Symmetry. $\varepsilon(H, E_1) > \varepsilon(H, E_2)$ *if and only if* $\varepsilon(H, \neg E_1) < \varepsilon(H, \neg E_2)$

Symmetry basically says that the better H is able to explain an explanandum E, the worse is its ability to explain E's complement $\neg E$. The next adequacy constraint goes back to Crupi and Tentori (2012):

Explanatory Justice (EJ). *If E' is statistically independent from E, H, and their conjunction $E \wedge H$, then:*

(i) if $\varepsilon(H, E) > 0$, then $\varepsilon(H, E \wedge E') < \varepsilon(H, E)$; and

(ii) if $\varepsilon(H, E) \leq 0$, then $\varepsilon(H, E \wedge E') = \varepsilon(H, E)$.

[11] See (Sprenger and Hartmann, 2019, p. 45).

[12] See (Sprenger and Hartmann, 2019, p. 56).

[13] Unlike $\varepsilon_{GMG}(H, E)$, $\varepsilon_{SS}(H, E)$ and $\varepsilon_{CT}(H, E)$ are not strictly identical to any confirmation measure that is supposed to measure the confirmation that E confers on H. But $\varepsilon_{SS}(H, E)$ and $\varepsilon_{CT}(H, E)$ are both structurally identical to certain well-known confirmation measures in the following way: $\varepsilon_{SS}(H, E)$ is the *Kemeny-Oppenheim Measure* with the roles of H and E reversed and $\varepsilon_{CT}(H, E)$ is the *Generalized Entailment Measure* with the roles of H and E reversed. See (Sprenger and Hartmann, 2019, p. 56) for a list of popular confirmation measures.

The intuition behind EJ(i) is that the amount of explanatory power that an explanation H has on an explanandum E, "cannot be extended 'for free"' (Crupi and Tentori, 2012, p. 370). Assume that H explains an explanandum E to a certain degree. Now imagine that we add some proposition E' to E, such that we obtain a logically stronger explanandum $E \wedge E'$. We have thereby increased the amount of information that we seek to explain. If neither H, nor E, nor the combination of $H \wedge E$ has any power to explain E', then the increased explanation-demand in the explanandum $E \wedge E'$ is uncompensated by the explanation-candidate H, which makes H a worse explanation for $E \wedge E'$ than it is for the less demanding explanandum E. The fact that an assassin has put a poison into Pete's tea seems to be a powerful explanation of Pete's death. But, to lend an example from (Schupbach and Sprenger, 2011), the fact that an assassin has put a poison into Pete's tea does not seem to be an equally powerful explanation of the following conjunction: Pete died and the mating season of the American green tree frog has just reached its height.

Now imagine that H has a negative explanatory power on an explanandum E. This means that learning H increases an agent's surprise about E rather than lowering it. Now, if H already increases the surprise about E, it should equally increase the surprise about a logically stronger statement $E \wedge E'$, as long as E, H, and $E \wedge H$ do not have any explanatory power on E'. Imagine that Pete does not die. Learning that an assassin has put a poison into Pete's tea has a negative explanatory power on our observation that Pete is alive. It makes Pete's vitality much more surprising. Adding the information, that the mating season of the American green tree frog has just reached its height, to the explanandum does not help here. The fact that an assassin has put a poison into Pete's tea is an equally bad explanation of the conjunction that Pete is alive and that the mating season of the American green tree frog has just reached its height. This, at least, is the intuition that is expressed by EJ(ii). Now, Sprenger and Hartmann (2019, cf. p. 197) prove the following representation theorem for ε_{CT}:

Representation Theorem for ε_{CT}. *All measures of explanatory power $\varepsilon(H, E)$ that satisfy GDM-EP, Symmetry, and EJ are ordinally equivalent to $\varepsilon_{CT}(H, E)$:*

$$\varepsilon_{CT} = \begin{cases} \dfrac{P(E|H) - P(E)}{1 - P(E)} & \text{if } P(E|H) \geq P(E), \\ \\ \dfrac{P(E|H) - P(E)}{P(E)} & \text{if } P(E|H) < P(E). \end{cases} \tag{8.5}$$

Finally, consider the following adequacy constraint by Sprenger and Hartmann (2019, cf. p. 197):

Independent Background Theories (IBT). *Suppose there is a theory T such that:*

$$P(H|E \wedge T) = P(H|E) \text{ and } P(H|\neg E \wedge T) = P(H|\neg E) \tag{8.6}$$

Then $\varepsilon(H, E) = \varepsilon_T(H, E)$, where ε_T refers to explanatory power calculated with respect to a probability distribution conditional on T.

Sprenger and Hartmann (2019) give the following motivation for IBT: "[I]f a theory T is irrelevant to the interaction between explanans H and explanandum E (and its negation $\neg E$), then conditionalizing on T does not affect the degree of explanatory power" (Sprenger and Hartmann, 2019, p. 198). Imagine that T represents the proposition that the mating season of the American green tree frog reaches its height in mid July, while our explanandum E is again the proposition that Pete dies, while H represents the proposition that an assassin has put a poison into Pete's tea. Clearly, we have $P(H|E,T) = P(H|E)$ and $P(H|\neg E,T) = P(H|\neg E)$. So, according to IBT, the explanatory power of the fact that an assassin has put a poison into Pete's tea on the fact that Pete has died, should not change when we actually come to learn that the mating season of the American green tree frog reaches its height in mid July. Now, Sprenger and Hartmann (2019, cf. p. 198) prove the following representation theorem:

Representation Theorem for ε_{SS}. *All measures of explanatory power $\varepsilon(H,E)$ that satisfy GDM-EP and IBT are ordinally equivalent to $\varepsilon_{SS}(H,E)$:*

$$\varepsilon_{SS}(H,E) = \frac{P(H|E) - P(H|\neg E)}{P(H|E) + P(H|\neg E)} \tag{8.7}$$

Table 8.1 summarizes which measure of explanatory power satisfies which adequacy constraint.

	GDM-EP	CV	Symmetry	EJ(i)	EJ(ii)	IBT
ε_{GMG}	yes	yes	no	no	yes	no
ε_{CT}	yes	no	yes	yes	yes	no
ε_{SS}	yes	no	yes	yes	no	yes

Table 8.1: Explanatory power measures and their respective adequacy constraints. Adapted from (Sprenger and Hartmann, 2019, p. 199).

There is a straightforward criticism that applies to all three measures. Imagine an event H that has a generative causal power on another event E. If we are ignorant about whether H and E are the case, then learning H would clearly increase the probability of E. Accordingly, ε_{GMG}, ε_{CT}, and ε_{SS} will all ascribe H a positive amount of explanatory power in respect to E. But in the same situation, learning E would also increase the probability of H. ε_{GMG}, ε_{CT}, and ε_{SS} will therefore also ascribe E a positive amount of explanatory power in respect to H. But this is clearly grotesque. An effect cannot explain its cause, which means that it cannot have any explanatory power in respect to its cause.

The proponents of the stated measures are well aware of the problem. In response to this inadequacy of the measures, they all accept that none of the three measures is able to identify an explanation. The identification of explanations is simply regarded as a separate problem that needs a separate solution.[14] The measures $\varepsilon_{GMG}(H,E)$, $\varepsilon_{CT}(H,E)$, and $\varepsilon_{SS}(H,E)$ can therefore only be used as measures of explanatory power, if H is already identified as an explanation of E. If H is no explanation of E, then $\varepsilon_{GMG}(H,E)$, $\varepsilon_{CT}(H,E)$, and $\varepsilon_{SS}(H,E)$ may very well take on non-zero values. But those values should simply not be interpreted as degrees of explanatory power.

[14] See, for example, (Schupbach and Sprenger, 2011, p. 107).

One might perceive this response as somewhat unsatisfactory, because it simply avoids to face a crucial and intricate challenge when it comes to evaluating explanations, namely the challenge of identifying explanations in the first place. But, when it comes to causal explanations, we can actually be content with the response, because the previous chapters of this dissertation are supposed to provide a solution to the problem of identifying explanations. So, whenever we want to evaluate the explanatory power of a given event H in respect to a given event E, we would simply have to stick to the following workflow: We first use our account of causal explanation to clarify whether H is a causal explanation of E. If it is not, H has no explanatory power in respect to E. If it is, then we can use $\varepsilon_{GMG}(H, E)$, $\varepsilon_{CT}(H, E)$, or $\varepsilon_{SS}(H, E)$ to determine the degree of explanatory power of H on E.

But then the question arises: Which of the three measures should we use to determine the explanatory power of a causal explanation H of E? The measures are not ordinally equivalent to each other, which means, that for a given explanandum E, each measure may very well lead us to prefer a different explanation. To decide which of the three measures is the most adequate, the presented adequacy constraints seem to provide a useful foundation. Apart from GDM-EP, the three measures differ in the constraints they satisfy. So, if we are able to constitue which adequacy constraints are really adequate, or at least, which are the most intuitive for a measure of explanatory power, then we are able to decide, which of the stated measures is superior as a measure of explanatory power. This is exactly the approach that has dominated the literature on explanatory power so far.[15] But I consider this debate to be futile for one very simple reason: I do not think that any of the three measures is adequate as a measure of explanatory power, at least when it comes to the explanatory power of causal explanations.

For starters, just as there was no reason to accept GDM for a measure of generative causal power, there is no good reason to believe that there is one specific function f of only the two arguments $P(E|H)$ and $P(E)$ that is able to determine the explanatory power of a causal explanation H on the explanandum E in any arbitrary situation. So, the first adequacy constraint GDM-EP already stands on very shaky grounds. But it is GDM-EP that all three measures satisfy and that is needed for all three representation-theorems by Sprenger and Hartmann (2019). If GDM-EP cannot be motivated by any intution, then the stated representation-theorems and the remaining adequacy constraints cannot provide any intuitive corroboration of either $\varepsilon_{GMG}(H, E)$, $\varepsilon_{CT}(H, E)$, or $\varepsilon_{SS}(H, E)$. Recognizing that GDM-EP does not have any firm intuitive underpinning already deprives any of the above adequacy constraints from its ability to support the adequacy of any of the three measures as a measure of explanatory power. But there is more than that. There is not just a lack of intuitive support for the adequacy of $\varepsilon_{GMG}(H, E)$, $\varepsilon_{CT}(H, E)$, or $\varepsilon_{SS}(H, E)$. There are also several reasons against the adequacy of any of the three measures as a measure of explanatory power.

[15] See, for example, (Schupbach and Sprenger, 2011), (Schupbach, 2011b), (Crupi and Tentori, 2012), (Crupi, 2012) (Cohen, 2015), (Cohen, 2016), (Cohen, 2018), (Sprenger and Hartmann, 2019).

8.4 Problems with the Bayesian Measures

8.4.1 The Explanatory Old Evidence Problem

Imagine that Lena comes home and she sees that her house has a broken window. Clearly, she will wonder why. Imagine that she gathers her kids, Suzy and Billy, since both were at home during the day, and she demands a causal explanation for the window being broken. Suzy and Billy do not waste any time and quickly confront her mom with several hypotheses to bury the truth: 'Perhaps a wind gust broke the window'. 'Or a squirrel threw a nut at it'. 'Maybe our neighbour listened to his music so loudly that the soundwaves broke the window'. 'Or maybe we have thrown some hand-sized stones at it'. It seems like a very desperate attempt to get away with their crime. But Suzy and Billy know what they are doing. They know that Lena always uses one of the three measures, ε_{GMG}, ε_{CT}, or ε_{SS}, to evaluate the explanatory power of a causal explanation. And since Lena already knows that the window is broken and therefore assigns a probability of 1 to the explanandum, she will not be able to recognize that one of the presented hypotheses has a significantly higher explanatory power than the others. $\varepsilon_{CT}(H, E)$ and $\varepsilon_{SS}(H, E)$ are undefined for every H, if $P(E) = 1$. And $\varepsilon_{GMG}(H, E)$ will assign every explanation H of E a power of 0, if $P(E) = 1$. Eva and Stern (2019) call the problem that Lena faces the *explanatory old evidence problem*: As soon as one aims to determine the explanatory power of an explanation for an explanandum, that has a probability of 1, the measures ε_{GMG}, ε_{CT}, and ε_{SS} are entirely useless.

The explanatory old evidence problem does not only arise in artificial or exceptional scenarios. Quite the contrary. It is rather typical to first learn about a certain event E and then, triggered by the new insight, to search for potential explanations for E and, subsequently, to evaluate those explanations. The fact that all three Bayesian measures are unable to cope with such a common and intuitively impeccable practice of assessing and evaluating explanations, is therefore very disquieting.

There are further related problems facing the Bayesian measures. For example, Lena might already know that Suzy and Billy threw hand-sized stones at the window, because she saw them doing so when she came home. But as soon as the explanation-candidate H has a probability of 1, conditionalizing on H cannot raise the probability of the explanandum E anymore. This is why $\varepsilon_{GMG}(H, E)$, $\varepsilon_{CT}(H, E)$, and $\varepsilon_{SS}(H, E)$ all ascribe an explanatory power of 0 to any H that is already believed to be true. But it seems intuitively very odd that, when looking for an explanation of the broken window, Lena ascribes an explanatory power of 0 to the fact that Suzy and Billy threw hand-sized stones at it, only because she saw them doing so.

As Glymour (2015) points out, we are also intuitively able to evaluate the explanatory power of explanation-candidates that we know to be false. For example, it seems reasonable to say that a hurricane would be a very powerful causal explanation of the broken window, even if we already know that there was no hurricane near Lena's house. But yet again, all three Bayesian measures are not applicable, since they are undefined for $P(H) = 0$.

One might hope to solve these problems by avoiding probability distributions that assign maximal or minimal values to explananda and explanation candidates. Instead of using a probability distribution that represents the actual current epistemic state of a given agent, one might, for example, try to refer to an hypothetical epistemic state, in which any maximal or minimal degree of belief concerning the explanation-candidate H or the explanandum E is retracted. But how exactly should such an hypothetical epistemic state look like? What degrees of beliefs should be assigned to H and E, if not the actual degrees of beliefs that the agent currently holds? Since, according to all three Bayesian measures, the exact probabilities of H and E affect the value of the explanatory power of H on E, this question needs an unambiguous answer. One might want to choose the probability distribution that represents the agents epistemic state shortly before she comes to adopt a maximal or minimal degree of belief about either H or E. But this would mean that the explanatory powers of distinct explanation-candidates, H_1 and H_2, are often not measured relative to one and the same probability distribution. It is therefore highly questionable whether the respective explanatory powers would be comparable. But comparing explanatory powers of competing explanations is the whole point.

8.4.2 Independence of Priors

Peter Lipton famously distinguished between the loveliness and the likeliness of an explanation. He writes:

> “It is important to distinguish two senses in which something may be the best of competing potential explanations. We may characterize it as the explanation that is most warranted: the ’likeliest’ or most probable explanation. On the other hand, we may characterize the best explanation as the one which would, if correct, be the most explanatory or provide the most understanding: the ’loveliest explanation’. The criteria of likeliness and loveliness may well pick out the same explanation in a particular competition, but they are clearly different sorts of standard. Likeliness speaks of truth; loveliness of potential understanding” (Lipton, 2004, p. 59).

Lipton not only considers the likeliness of an explanation H and the loveliness of H as an explanation of E to be two different sorts of standards, he also points out that a change in the likeliness of H does not influence the loveliness of H as an explanation of E.[16] He notes: “More recently, with the advent of special relativity and the new data that support it, Newtonian mechanics has become less likely, but it remains as lovely an explanation of the old data as it ever was” (Lipton, 2004, p. 60). To illustrate this intuition further, consider again Lena and the broken window. A squirrel throwing a nut at the window amounts to an intuitively very weak explanation of the window breaking. It also seems to be a rather unlikely event. But imagine now that Lena has just read in the local newspaper, that nut throwing squirrels have lately been observed all around town. The proposition that a squirrel threw a nut at her window just became much more likely for Lena. But as an explanation of the window breaking it remains just as weak as it has been before. The weakness of the explanation is intuitively grounded in the weak causal relationship between a squirrel’s nut throw and the window breaking. It has

[16] Notice though that the loveliness of H as an explanation of E may very well influence the probability that we assign to H. As Lipton (2004, cf. p. 60) points out, this is just the idea of inference to the best (loveliest) explanation: If we come to learn that E is the case and if H is the loveliest explanation of E, then this propels us to also increase our degree of belief in H.

nothing to do with the probability of the explanation candidate itself.[17]

Clearly, there is a sense in which a less plausible explanation is a worse explanation. Imagine that Suzy and Billy claim: 'An alien appeared in front of our house and it fired its laser gun, which broke the window'. Even though firing a laser gun at a window has probably a very high causal power on the window breaking, the explanation is intuitively bad, for the simple reason that it is highly implausible. But the badness of the alien-explanation differs from the badness of the squirrel-explanation. If an alien would actually appear and fire its laser gun at the window, then this would amount to a very powerful explanation of the window breaking. But if a squirrel would actually throw a nut at the window, it would still amount to a weak explanation of the window breaking. Separating the power of a potential explanation from its likeliness means to separate two dimensions of being a good or a bad explanation.[18]

Schupbach (2011b) seems to be sympathetic to the distinction between the likeliness of an explanation candidate H and the power of H to explain a given explanandum E. He also seems to agree that the first, the likeliness of H, does not influence the second, the power of H to explain E. While arguing for ε_{SS} as the adequate measure of explanatory power, Schupbach notes: "[T]he extent to which an explanatory hypothesis alleviates the surprising nature of some explanandum does not depend on considerations of how likely that hypothesis is in and of itself" (Schupbach, 2011b, p. 42). To illustrate this idea, he gives the following example:

> "[D]ehydration and cyanide poisoning may be (approximately) equally powerful explanations of symptoms of dizziness and confusion insofar as they both make such symptoms less surprising to the (approximately) same degree. And this is true despite the fact that dehydration is typically by far the more plausible explanans" (Schupbach, 2011b, p. 43).

Schupbach then even formulates an adequacy constraint for any measure $\varepsilon(H, E)$ of explanatory power, according to which the values of $\varepsilon(H, E)$ should not depend upon the values of $P(H)$.[19] He subsequently suggests that ε_{SS} satisfies this constraint. But this is misleading. Even though $P(H)$ does not explicitly appear in the measure $\varepsilon_{SS}(H, E)$ and even though it is, at least in principle, possible to determine the value of $\varepsilon_{SS}(H, E)$ without knowing the value of $P(H)$,[20] the value of $\varepsilon_{SS}(H, E)$ is not independent from changes to the value of $P(H)$. Actually, all three Bayesian measures, $\varepsilon_{GMG}(H, E)$, $\varepsilon_{CT}(H, E)$, and $\varepsilon_{SS}(H, E)$, are highly sensitive to changes of the prior probability of H.

Consider as an example a probabilistic version of the disjunctive forest fire scenario. Figure 8.1 shows a probabillistic SEM representing the scenario, in which A represents whether there is

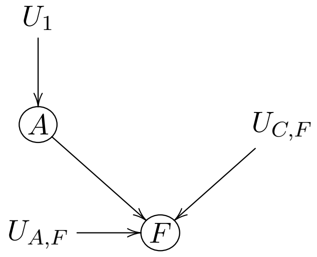
[17] Notice that this is basically a generalization of the intuition expressed in the last section: There we have argued that whether an explanation-candidate H is assumed to be true or false does not affect the value of its explanatory power on a given explanandum E. We now argue: The degree of belief in whether H is true does not affect the value of its explanatory power on E. Both claims can be summarized by saying that H's ability to render E less surprising, does not depend on whether or on how strongly we belief that H is true.

[18] We should actually differentiate between three different concepts here. There is the likeliness of the explanation candidate H. Then there is the explanatory power that H has on a given explanandum E. And finally, there is the likeliness of whether H is an explanation of E.

[19] See (Schupbach, 2011b, p. 43).

[20] This is possible, as long as we can epistemically access the values of $P(H|E)$ and $P(H|\neg E)$ without knowing the value of $P(H)$.

an arsonist dropping a match in the forest and $U_{C,F}$ represents whether there is any other cause that causally produces $F = 1$.



- $A := U_1$
- $F := (A \wedge U_{A,F}) \vee U_{C,F}$

Figure 8.1: A probabilistic version of the disjunctive forest fire scenario.

Let us say that $A = 1$ has a generative causal power of 0.9 on $F = 1$, which means that $\mathcal{P}(U_{A,F} = 1) = 0.9$, and let us assume that $\mathcal{P}(U_{C,F} = 1) = 0.08$ and $\mathcal{P}(U_1 = 1) = 0.9$. This gives us $\mathcal{P}(F = 1) \approx 0.825$. $A = 1$ is an extensive causal explanation of $F = 1$.[21] According to their defenders, we can therefore apply the Bayesian measures to determine the explanatory power of $A = 1$ on $F = 1$. Here are the probability values that we need for the application of ε_{GMG}, ε_{CT}, and ε_{SS}:[22]

- $\mathcal{P}(F) \approx 0.825$
- $\mathcal{P}(F|A) \approx 0.908$
- $\mathcal{P}(A|F) \approx 0.817$
- $\mathcal{P}(\neg F) \approx 0.175$
- $\mathcal{P}(\neg F|A) \approx 0.092$
- $\mathcal{P}(A|\neg F) \approx 0.473$

We now have:

$$\varepsilon_{SS}(A, F) = \frac{\mathcal{P}(A|F) - \mathcal{P}(A|\neg F)}{\mathcal{P}(A|F) + \mathcal{P}(A|\neg F)} \approx \frac{0.817 - 0.473}{0.817 + 0.473} \approx 0.27 \tag{8.8}$$

$$\varepsilon_{CT}(A, F) = \frac{\mathcal{P}(F|A) - \mathcal{P}(F)}{1 - \mathcal{P}(F)} \approx \frac{0.908 - 0.825}{1 - 0.825} \approx 0.47 \tag{8.9}$$

$$\varepsilon_{GMG}(A, F) = ln\frac{\mathcal{P}(F|A)}{\mathcal{P}(F)} \approx ln\frac{0.908}{0.825} \approx 0.1 \tag{8.10}$$

Let us now imagine that the prior probability of A is changed to $\mathcal{P}(A) = 0.1$, while we still have $\mathcal{P}(U_{A,F}) = 0.9$ and $\mathcal{P}(U_{C,F}) = 0.08$. We then have the following probabilities:

- $\mathcal{P}(F) \approx 0.1628$
- $\mathcal{P}(F|A) \approx 0.908$
- $\mathcal{P}(A|F) \approx 0.558$
- $\mathcal{P}(\neg F) \approx 0.837$
- $\mathcal{P}(\neg F|A) \approx 0.092$
- $\mathcal{P}(A|\neg F) \approx 0.011$

[21] To see that this is the case, notice that there is a non-zero probability of $A = 1$ being a strong actual cause of $F = 1$. This makes $A = 1$ an explicitly complete explanation of $F = 1$. Since $A = 1$ is also a maximal potential strong actual cause of $F = 1$ relative to the given causal model, $A = 1$ is an extensive causal explanation according to rule (a) of our definition of extensive causal explanations.

[22] For a simpler notation, I will just write A for $A = 1$, $\neg A$ for $A = 0$, F for $F = 1$, and $\neg F$ for $F = 0$ in the example at hand.

We then have:

$$\varepsilon_{SS}(A,F) = \frac{\mathcal{P}(A|F) - \mathcal{P}(A|\neg F)}{\mathcal{P}(A|F) + \mathcal{P}(A|\neg F)} \approx \frac{0.558 - 0.011}{0.558 + 0.011} \approx 0.96 \tag{8.11}$$

$$\varepsilon_{CT}(A,F) = \frac{\mathcal{P}(F|A) - \mathcal{P}(F)}{1 - \mathcal{P}(F)} \approx \frac{0.908 - 0.1628}{1 - 0.1628} \approx 0.89 \tag{8.12}$$

$$\varepsilon_{GMG}(A,F) = ln\frac{\mathcal{P}(F|A)}{\mathcal{P}(F)} \approx ln\frac{0.908}{0.1628} \approx 1.72 \tag{8.13}$$

This illustrates that, ceteris paribus, a decrease in the prior probability of the explanation H leads to an increase of H's power to explain a given explanandum, according to all three Bayesian measures. In the given example, this means the following: The less you believe that there was an arsonist kindling in the forest, the more powerful this becomes as a potential explanation of a forestfire. This is not just highly counterintuitive, it also shows that all three Bayesian measures act in opposition to the idea that the likeliness of an explanation candidate should not influence its power as an explanation of a given explanandum, which is an idea that Schupbach himself explicitly endorsed.

If we measure the explanatory power of an explanation H on a given explanandum E by either $\varepsilon_{SS}(H,E)$ or $\varepsilon_{GMG}(H,E)$, then there is another factor that influences H's power to explain E: The likeliness (and the effective causal influence) of alternative causes of E. Consider the previous example with $\mathcal{P}(U_1 = 1) = 0.1$ and $\mathcal{P}(U_{A,F} = 1) = 0.9$. But this time, imagine that $\mathcal{P}(U_{C,F} = 1) = 0.7$, which means that the probability of there being some alternative cause, besides A, that causally produces F is considered to be significantly higher than in the previous scenario, where we assumed that $\mathcal{P}(U_{C,F} = 1) = 0.08$. We then have the following probabilities:

- $\mathcal{P}(F) \approx 0.727$
- $\mathcal{P}(F|A) \approx 0.97$
- $\mathcal{P}(A|F) \approx 0.133$
- $\mathcal{P}(\neg F) \approx 0.273$
- $\mathcal{P}(\neg F|A) \approx 0.03$
- $\mathcal{P}(A|\neg F) \approx 0.011$

This gives us:

$$\varepsilon_{SS}(A,F) = \frac{\mathcal{P}(A|F) - \mathcal{P}(A|\neg F)}{\mathcal{P}(A|F) + \mathcal{P}(A|\neg F)} = \frac{0.133 - 0.011}{0.133 + 0.011} \approx 0.85 \tag{8.14}$$

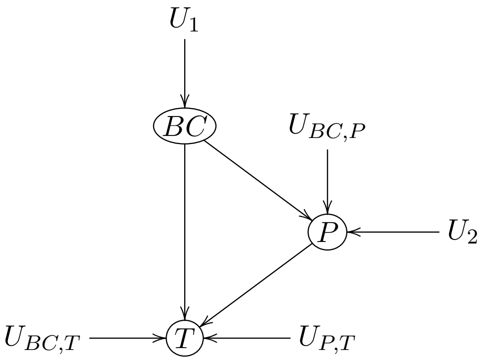
$$\varepsilon_{CT}(A,F) = \frac{\mathcal{P}(F|A) - \mathcal{P}(F)}{1 - \mathcal{P}(F)} = \frac{0.97 - 0.727}{1 - 0.727} \approx 0.89 \tag{8.15}$$

$$\varepsilon_{GMG}(A,F) = ln\frac{\mathcal{P}(F|A)}{\mathcal{P}(F)} = ln\frac{0.97}{0.727} \approx 0.29 \tag{8.16}$$

The results illustrate that for $\varepsilon_{GMG}(A,F)$ and $\varepsilon_{SS}(A,F)$ the following holds: Ceteris paribus, the more you believe that some other cause, besides A, causally produced the forest fire, the less powerful becomes A as an explanation of the forest fire. Yet again, I consider this result to be highly counterintuitive. Intuitively, the explanatory power of A on F seems to be independent of the probability of alternative causes of F. Imagine that some more or less reliable witness tells us that there has been a lightning strike, which is clearly an alternative potential causal producer of a forest fire. Learning about the lightning strike provides us with an additional explanation candidate for a forest fire, an explanation candidate to which we now assign a rather high probability. But, intuitively, all this does not influence the explanatory power of A on F. A kindling arsonist in the forest is just as powerful an explanation of a forest fire as it has been before we increased our degree of belief in a lightning strike.

8.4.3 Confounding

Glymour (2015) draws attention to another weakness of the Bayesian measures of explanatory power, namely the problem of confounding. Consider the following scenario, which is based on an example by Hesslow (1976). An agent α believes to a certain degree that Lena takes birth control pills ($BC = 1$), which, by some causal mechanism, increases the risk of her getting thrombosis ($T = 1$). But it also lowers the risk of her becoming pregnant ($P = 1$). Being pregnant, however, increases the risk of getting thrombosis. The scenario can be represented by the causal model in figure 8.2.



- $BC := U_1$
- $P := U_2 \wedge \neg(BC \wedge U_{BC,P})$
- $T := (BC \wedge U_{BC,T}) \vee (P \wedge U_{P,T})$

Figure 8.2: Causal model of the thrombosis scenario.

In case Lena gets thrombosis, her being pregnant would be an extensive causal explanation for it.[23] Now, imagine that we want to determine the explanatory power of $P = 1$ on $T = 1$ with one of the Bayesian measures. The values of $\varepsilon_{GMG}(P = 1, T = 1)$ and $\varepsilon_{CT}(P = 1, T = 1)$ crucially rely on the value of $\mathcal{P}(T = 1|P = 1)$. But conditionalizing on $P = 1$ does not only have a positive influence on the probability of $T = 1$ due to the direct causal path from $P = 1$ to $T = 1$. Conditionalizing on $P = 1$ also lowers the probability of $BC = 1$, which, due to the positive probabilistic influence of $BC = 1$ on $T = 1$, leads to a negative influence on the probability of $T = 1$. It may even be the case, that these two paths of influencing the probability of $T = 1$ completely cancel each other out. In that case $\mathcal{P}(T = 1|P = 1) = \mathcal{P}(T = 1)$, which means that $\varepsilon_{GMG}(P = 1, T = 1) = \varepsilon_{CT}(P = 1, T = 1) = 0$.[24] But $P = 1$ is intuitively still a

[23] Here again, $P = 1$ is a maximal potential strong actual cause of $T = 1$ relative to the given causal model.

[24] The confounding is just as bad, though, if both paths of influence on the probability of $T = 1$ do not completely cancel each other out. The confounding still distorts the value of explanatory power.

powerful causal explanation of $T = 1$. Believing in $P = 1$ only lowers the degree of belief in $BC = 1$, which is an alternative causal explanation of $T = 1$. But this reduced belief in an alternative explanation of $T = 1$ should not reduce the power of $P = 1$ to explain $T = 1$.

The measure ε_{SS} does not perform any better in this scenario. The value of $\varepsilon_{SS}(P = 1, T = 1)$ crucially relies on the value of $\mathcal{P}(P = 1|T = 1)$. But conditionalizing on $T = 1$ also influences the probability of $P = 1$ by two distinct paths. It increases the probability of $P = 1$ due to the direct causal path from $P = 1$ to $T = 1$. But conditionalizing on $T = 1$ also increases the probability of $BC = 1$ due to the direct causal path from $BC = 1$ to $T = 1$, which in turn leads to a decrease of the probability of $P = 1$. So, here as well, confounding distorts the value of $\varepsilon_{SS}(P = 1, T = 1)$.

One might hope to avoid confounding in the determination of $\varepsilon_{GMG}(P = 1, T = 1)$, $\varepsilon_{CT}(P = 1, T = 1)$, or $\varepsilon_{SS}(P = 1, T = 1)$ by simply conditioning on the common cause $BC = 1$. But, as Glymour points out: "conditioning on different values of the common causes will give different values to the measures of explanatory power" (Glymour, 2015, p. 596). Conditioning on $BC = 1$ lowers the probability of $P = 1$ and increases the probability of $T = 1$. Conditioning on $BC = 0$ has the opposite effect. And as we have seen, both the probability of $P = 1$ and the probability of $T = 1$ highly influence the results of the Bayesian explanatory power measures. It therefore remains unclear, on which value of the common cause one should conditionalize for the determination of explanatory power. Just opting for one of the possible options is an arbitrary and unjustified decision, which makes the resulting value of explanatory power equally arbitrary and unjustified.

8.4.4 Actual Surprise Reduction vs. Surprise Reduction Ability

Let us briefly take stock. So far, we have presented three deficiencies of the Bayesian measures as measures of explanatory power. All three deficiencies clearly illustrate that $\varepsilon_{GMG}(H, E)$, $\varepsilon_{CT}(H, E)$, and $\varepsilon_{SS}(H, E)$ do not really measure H's "ability to decrease the degree to which we find the explanandum [E] surprising" (Schupbach and Sprenger, 2011, p. 108). Instead they rather seem to be geared to measure the surprise reduction concerning the explanandum E that a given agent α actually experiences in a given situation after assuming that H is true. This would at least explain why every time that an agent α already fully believes in the truth of E or H, the Bayesian measures are either undefined or yield a value of 0, because in those cases, α would indeed not experience any actual surprise reduction concerning E, if she would assume that H is true. It would also explain why changes in the likeliness of H, or in the likeliness of alternative explanations of E, influence the results of the Bayesian measures, since a higher likeliness of H, or a higher likeliness of alternative explanations of E, leads to a higher probability of E. This in turn decreases the amount of surprise reduction that α would actually experience after conditionalizing on H. But changes in the likeliness of H, E, or alternative causal explanations of E, do not influence the *ability* of H "to decrease the degree to which we find the explanandum [E] surprising" (Schupbach and Sprenger, 2011, p. 108), because this ability is an intrinsic capacity of H, which is unaffected by changes of the likeliness of H, E, or alternative explanations of E. As Schupbach and Sprenger's (2011) informal characterization already suggests, it is this ability, this intrinsic capacity of H, that we intuitively identify with

H's explanatory power on E. The Bayesian measures, $\varepsilon_{GMG}(H, E)$, $\varepsilon_{CT}(H, E)$, and $\varepsilon_{SS}(H, E)$, are unable to quantify this intrinsic capacity of H. This is why they so often yield results that clearly diverge from our intuitions about explanatory power. And there is yet another example that nicely demonstrates this divergence.

8.4.5 Flipping Coins

Cohen (2016) presents the following example: Alice has two coins that look identical. But she knows that one coin is fair and the other is biased. She knows that the fair coin has a 50% chance of landing heads, if tossed, while the biased coin has a 70% chance of landing heads, if tossed. Now, Alice takes one of the two coins and starts tossing. She tosses k times with the same coin and in every single instance of the k tosses the coin lands on heads. Cohen applies all three Bayesian measures of explanatory power to determine how well the hypothesis $H =$ 'The coin is biased with a bias of 70% towards heads', explains the explanandum $E_k =$ "All k coin-tosses landed on heads". The results are summarized in table 8.2.

Number of flips (k)	$\varepsilon_{SS}(H, E_k)$	$\varepsilon_{CT}(H, E_k)$	$\varepsilon_{GMG}(H, E_k)$
1	0.217	0.250	0.154
2	0.241	0.190	0.281
3	0.262	0.142	0.382
4	0.279	0.105	0.462
5	0.292	0.076	0.523
10	0.324	0.014	0.659
20	0.333	0.000	0.692
30	0.333	0.000	0.693

Table 8.2: Values of explanatory power in the coin-tossing example. Adapted from (Cohen, 2016, p. 1084).

Interestingly, Cohen introduces the example in order to argue that ε_{SS} and ε_{GMG} are more adequate measures of explanatory power than ε_{CT}. But in fact, the example nicely illustrates that all three Bayesian measures are inadequate for representing what we intuitively understand under explanatory power.

Consider the case with $k = 2$. All three measures ascribe a rather low explanatory power of H on E_2. But if Alice throws a coin twice and the coin lands heads twice, then the information that the coin has a 70% bias towards heads appears to be a fairly strong explanation of the outcome. If we instead have an outcome with, for example, 10 heads in 10 tosses, then the information that the coin has a 70% bias towards heads is still somewhat helpful in explaining the outcome, but much less so than in the scenario with 2 heads in 2 tosses. The reason is that 10 heads in 10 tosses is quite unlikely, even for a coin with a 70% bias towards heads ($P(E_{10}|H) = 0.7^{10} \approx 0.028$). Learning or assuming that the coin has a 70% bias towards heads can therefore not significantly lower our surprise about the outcome of 10 heads in 10 tosses. 2 heads in 2 tosses, on the other hand, is an expectable outcome, given that the coin has a 70% bias towards heads ($P(E_2|H) = 0.49$). In general, it therefore seems intuitive that for increasing k, H becomes a less and less powerful explanation for E_k. But only ε_{CT} accords with this intuition. ε_{SS} and ε_{GMG} display just the opposite behaviour. According to these two measures, the explanatory power of H grows for increasing K's, until it converges to a fixed value: 0.333 for ε_{SS} and 0.693 for ε_{GMG}. But this is highly counterintuitive. Just imagine that Alice tosses the coin 1000 times and the coin lands heads each time. This is nearly inexplicable for a coin with a 70% bias towards heads. Learning that the coin has a 70% bias towards heads will still leave us, not only surprised, but utterly confused and shocked about the outcome ($P(E_{1000}|H) = 1.25^{-155}$). But both ε_{SS} and ε_{GMG} consider H to be a more powerful explanation of E_{1000} than it is of E_2. This is grotesque.

Cohen defends these results by stating: "[A] measure of explanatory power is supposed to reflect the situation that happens when the background information holds, not to second guess it" (Cohen, 2016, p. 1085). Remember, that Alice presupposed that there are only two alternatives: Either the coin is fair or the coin has a 70% bias towards heads. Clearly, the second hypothesis does a better job than the first in explaining E_2 as well as in explaining E_{1000}. But still, even with the background information that there are only these two options, Alice should be able to recognize that both options are extremely bad explanations for E_k with k being large.

ε_{CT} performs significantly better in this example. As pointed out, it accords with the intuition that the explanatory power of H on E_k should decrease for increasing k. Still, ε_{CT} does not deliver intuitively persuasive results. As already mentioned, it ascribes unintuitively low values to the explanatory power of H on E_k for small values of k, like $k = 1$ or $k = 2$, and it continues to do so for slightly larger values of k. Just consider the case with $k = 5$. Alice tosses the coin five times and it lands heads every single time. The information that the coin has a 70% bias towards heads, though far from being a perfect explanation, is still able to alleviate some of our surprise about E_5. After all, we have $P(E_5|H) = 0.168$. But ε_{CT} ascribes H a vanishingly low explanatory power of 0.076 on E_5.

8.5 Other Proposals for Measuring Explanatory Power

8.5.1 Gärdenfors' Measure of Explanatory Power

I will just briefly point to some further proposals that have been made to measure explanatory power. Gärdenfors (1988) puts forward the formula $P(E|H) - P(E)$ to measure the explanatory power of H on E, while Chajewska and Halpern (1997) propose the following formula:

$$\frac{P(E|H)}{P(E)} \tag{8.17}$$

In both proposals, the probability distribution P is assumed to be a distribution, in which the full belief in E is contracted, which is an attempt to avoid the explanatory old evidence problem. But it does not need much effort to see that both proposals suffer from the same problems as the three Bayesian measures, even if we assume that the explanatory old evidence problem is solved.[25] According to both measures, the explanatory power of an explanation candidate H on a given explanandum E is influenced by the likeliness of H itself and by the prevelance and strength of other causes of E. Both measures face the same problems of confounding and they

[25] Chajewska and Halpern's measure is actually ordinally equivalent to ε_{GMG}.

give similarly strange results in Cohen's coin-tossing example.

8.5.2 Eva and Stern's Causal Explanatory Power

To tackle the explanatory old evidence problem Eva and Stern (2019) suggest some adjustments to the Bayesian measures.[26] First and foremost, they suggest that the probability distribution, that is employed in the Bayesian measures, should not be understood as representing an agent's actual degrees of belief about whether certain events did or do happen. Instead, the distribution should correspond "to what the causal details of the system give us reason to expect" (Eva and Stern, 2019, p. 1035) and it should be "consistent with the causal relationships (or structural equations) that govern the causal system" (Eva and Stern, 2019, p. 1034). Eva and Stern call such a probability distribution a *causal distribution*.[27] By only employing causal distributions, that are formulated in the framework of causal models, for the Bayesian measures of explanatory power, Eva and Stern (2019) limit their account of explanatory power to causal explanations only, while the original Bayesian measures of explanatory power can in principle be applied to any kind of explanation. This is why Eva and Stern call the quantity that their approach aims to measure *causal explanatory power*.[28]

Notice, though, that the constraints for being a causal distribution do not determine a unique probability distribution. In a causal model, there can be several causal distributions that differ in their probability assignments to values of the endogenous variables, if they ascribe different prior probabilities to the values of the exogenous variables. So, how should these prior probabilities over the exogenous variables be determined for yielding unambiguous values of explanatory power? Here is the answer that Eva and Stern propose in a footnote:

> "The reader may harbour doubts about our capacity to determine the unconditional probability estimates of exogenous variables since these are not supplied by the structural equations. We are sympathetic to these concerns and are correspondingly ecumenical about how these probabilities should be set. In the absence of knowledge of long-run frequencies, one might justifiably use tools in the objective Bayesian toolbox to determine these estimates – for example, principles of indifference" (Eva and Stern, 2019, p. 1034).

A tool from the objective Bayesian toolbox might indeed provide us with a unique causal distribution that can be employed in the Bayesian measures of explanatory power. But we should keep in mind that the result of a Bayesian measure of explanatory power is strongly influenced by the prior probability of the explanation and the prior probabilities of alternative causes of the explanandum. Which tool from the objective Bayesian toolbox we choose to yield our causal

[26] Eva and Stern (2019) do not explicitly specify which of the Bayesian measures they favour. They are more generally concerned with the basic sentiment of all three measures, which is that, ceteris paribus, the degree of explanatory power increases with an increasing degree of statistical relevance between the explanation and the explanandum (Eva and Stern, 2019, cf. 1032-33).

[27] In the framework of probabilistic causal models, the preliminary distribution that results from step (1) in PROBAC and PROSAC is a causal distribution. It is the distribution in which the probabilities over the error-terms still correspond with the respective values of intrinsic causal power.

[28] As we will see in the next section, my own account of explanatory power will likewise be limited to causal explanations only.

distribution has therefore a very strong influence on the values of explanatory power that we obtain. But, more crucially, allowing only causal distributions for the Bayesian measures of explanatory power does not necessarily avoid the explanatory old evidence problem and the related problems discussed in section 8.4.1. A causal distribution might very well assign a probability of 1 or 0 to an explanation candidate or to other causes of the explanandum of interest. If any of these causes has a maximal generative causal power on the explanandum, then a causal distribution can also assign a probability of 1 to the explanandum. It is therefore not quite clear, why the restriction to causal distributions in the Baysian measures of explanatory power should solve the problems discussed in section 8.4.1.

But there are two additional adjustments to the Bayesian measures that Eva and Stern (2019) propose. One of these additional adjustments is a partial departure from their first proposal to employ only causal distributions in the Bayesian measures. In the course of their paper, Eva and Stern (2019) argue that, while not all of our credences about the actual occurrences of events are relevant for adequately measuring causal explanatory power, some of them are. They therefore put forward the following guideline for determining causal explanatory power:

> (1) "First, represent the causal system in which the explanans and explanandum are embedded and work out the causal distribution for that system, according to our current best knowledge.
>
> (2) Second, update on all your background knowledge (excluding only the explanans and the explanandum themselves) regarding the causal system by intervening to set the relevant variables to their known values.
>
> (3) Calculate the explanatory power that intervening to make the explanans true exerts over the explanandum using the [Bayesian] measure ε, relative to the updated causal distribution" (Eva and Stern, 2019, p. 1044-45).

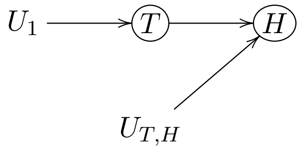
So, Eva and Stern employ interventions to reintroduce a large amount of an agent's credences about the actual values of endogenous variables into the causal distribution. They only exclude the credences about the explanation candidate and the explanandum, in the hope that this will avoid the explanatory old evidence problem. But this hope is clearly not met. As soon as we incoporate into the causal distribution an agents knowedge about the occurrence of some alternative cause of the explanandum that has a maximal generative causal power on it, the explanandum will have a probability of 1 in the causal distribution. And as soon as we incoporate into the causal distribution an agents knowedge about the presence of some cause of the explanation candidate that has a maximal generative causal power on it, the explanation candidate will have a probability of 1 in the causal distribution. The explanatory old evidence problem and the related problems that we have discussed in section 8.4.1 are therefore not banished. Furthermore, according to Eva and Stern's account, the background knowledge about the actual occurrences of events strongly influences the probabilities assigned to the explanation candidate and the explanandum. Since the Bayesian measures are strongly influenced by these probability values, the given background knowledge about the actual occurrences of events strongly influences the values of explanatory power. This leads to similarly unintuitive results for Eva and Stern's measure of causal explanatory power, as the ones that we have presented for the Bayesian measures in section 8.4.2.

There is yet one additional adjustment to the Bayesian measures that Eva and Stern (2019) propose. While the Bayesian measures all employ conditional probabilities, like $P(E|H)$, in which we describe the probability of the explanandum E after conditionalizing on a certain event, like H, Eva and Stern propose to use intervention counterfactuals, like $P(E|do(H))$, in which we describe the probability of E after realizing a certain event, like H, by an intervention. This adjustment is motivated by the hope that the resulting measures of causal explanatory power can not only determine the explanatory power of already known causal explanations of a given explanandum, but even to detect causal explanations of the given explanandum. This hope does not realize, though. There may very well be causal explanations of a given explanandum E whose realization by an intervention does not make any difference to the probability of E or whose realization even render E less probable. The mafia scenario in section 2.3.2, in which one potential causal producer of the explanandum has an even more powerful backup cause, is a typical example for this. But Eva and Stern's proposal to employ intervention counterfactuals instead of conditional probabilities has the positive effect of avoiding problems of confounding, which, as we have seen in section 8.4.3, trouble the original Bayesian measures. However, this does not save Eva and Stern's account of causal explanatory power, since using intervention counterfactuals instead of conditional probabilities does not help with the explanatory old evidence problem or with the deficiency that values of causal explanatory power highly depend on the background beliefs about which events in the scenario under consideration actually occurred.

8.5.3 Halpern and Pearl's Measure of Explanatory Power

In Chapter 1, we have already discussed Halpern and Pearl's proposal for measuring the degree of completeness of a partial explanation $\vec{X} = \vec{x}$ of ϕ, namely: $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi}|\vec{X} = \vec{x})$. I have argued that, given our definitions of explanation, $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi}|\vec{X} = \vec{x})$ is inept for measuring the degree of completeness of a partial explanation, which instead should be measured by $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi})$. But Halpern and Pearl (2005b) not only put forward a measure for measuring the degree of completeness of a partial explanation. They additionally advance another, though quite similar, measure for measuring, what they call, the explanatory power of explanations: $\mathcal{P}^-(\mathcal{K}_{\vec{X}=\vec{x},\phi}|\vec{X} = \vec{x})$. This explanatory power measure only differs from the completeness measure $\mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi}|\vec{X} = \vec{x})$ by the employed probability distribution $\mathcal{P}^-$, which "intuitively represents the agent's 'pre-observation' probability, that is, the agent's prior probability before the explanandum ϕ is observed or discovered" (Halpern and Pearl, 2005b, p. 905). Accordingly, Halpern and Pearl's measure of explanatory power is confronted with a similar challenge as the Bayesian measures, since it has to make precise how the "pre-observation" probabilities should be determined. But even if we us assume that we have a viable solution to this problem, Halpern and Pearl's proposal does still not amount to an adequate measure of explanatory power.

To see why, consider the following example: Lena tosses a coin that has a 90% bias towards tails. We can represent the scenario with the causal model $\mathcal{M}^C$ as shown in figure 8.3. The bivalent variable T represents whether Lena tosses the coin ($T = 1$ means that Lena tosses the coin; $T = 0$ means that Lena does not toss the coin) and the bivalent variable H represents whether the coin lands on heads ($H = 1$ means that the coin lands on heads; $H = 0$ means that the coin lands on tails). The error-term $U_{T,H}$ represents whether the coin toss ($T = 1$) causally produces the coin to land on heads ($H = 1$), given that $T = 1$ occurs. With the given bias, we therefore have: $\mathcal{P}(U_{T,H} = 1) = 0.1$.



- $T := U_1$
- $H := T \wedge U_{T,H}$

Figure 8.3: $\mathcal{M}^C$ - the coin-tossing example.

We assumed to know that Lena tosses the coin, which gives us $\mathcal{P}(U_1 = 1) = 1$. Imagine that we did not yet check whether the coin landed on heads or on tails. We therefore have: $\mathcal{P}(H = 1) = 0.1$.

Now, $T = 1$ is an extensive causal explanation of $H = 1$ in $(\mathcal{M}^C, \mathcal{P})$, since $\mathcal{P}(T = 1 \rightarrowtail^{\mathcal{M}^C} H = 1) > 0$ and since $T = 1$ is a maximal potential strong actual cause of $H = 1$ relative to $\mathcal{M}^C$. So, let us apply Halpern and Pearl's measure to determine the explanatory power of $T = 1$ on $H = 1$. Since $\mathcal{P}$ represents an epistemic state that does include knowledge about the actual value of H, $\mathcal{P}$ satisfies the condition of being a 'pre-observation' probability $\mathcal{P}^-$. Now, remember that $\mathcal{K}$ is the set of all the contexts for a given causal model $\mathcal{M}$, to which a given agent α assigns a non-zero degree of belief, while $\mathcal{K}_{\vec{X}=\vec{x},\phi}$ is the largest subset of $\mathcal{K}$, relative to which $\vec{X} = \vec{x}$ is an explanation of ϕ. Since $T = 1$ is an explanation of $H = 1$ relative to $\mathcal{K}$, we have $\mathcal{K}_{\vec{X}=\vec{x},\phi} = \mathcal{K}$, and accordingly: $\mathcal{P}^-(\mathcal{K}_{\vec{X}=\vec{x},\phi}|\vec{X} = \vec{x}) = \mathcal{P}(\mathcal{K}_{\vec{X}=\vec{x},\phi}|\vec{X} = \vec{x}) = \mathcal{P}(\mathcal{K}|\vec{X} = \vec{x}) = 1$.

So, according to Halpern and Pearl's measure, the explanatory power of $T = 1$ on $H = 1$ is maximal. This is clearly a highly counterintuitive result. Intuitively, tossing a coin with a 90% bias towards tails is an extremely weak explanation of the coin landing on heads.[29] Notice that we could even change the value of $\mathcal{P}(U_{T,H} = 1)$, which represents the bias of the coin, to any value $b > 0$, and we would still obtain the very same result that $\mathcal{P}^-(\mathcal{K}_{\vec{X}=\vec{x},\phi}|\vec{X} = \vec{x}) = 1$. So, according to Halpern and Pearl's measure, tossing a coin with a 90% bias towards tails and tossing a coin with a 99% bias towards heads are equally powerful explanations of the coin landing on heads.

Here is what Halpern and Pearl's measure $\mathcal{P}^-(\mathcal{K}_{\vec{X}=\vec{x},\phi}|\vec{X} = \vec{x})$ actually measures: It gives us the 'pre-observation' probability of $\vec{X} = \vec{x}$ being an explanation of ϕ, given that $\vec{X} = \vec{x}$ is the case. The previous example has clearly illustrated, that this is not the same as the intrinsic ability of $\vec{X} = \vec{x}$ to reduce an agent's surprise about ϕ, which is the explanatory power of $\vec{X} = \vec{x}$ on ϕ.

[29] One might want to object, that $T = 1$ is the only possible explanation for $H = 1$ in the given causal model, which makes it as good as it can get. I have two replies to this objection. First, we could easily amend the causal model and introduce alternative potential explanations of $H = 1$ and Halpern and Pearl's measure would still ascribe $T = 1$ a maximal explanatory power on $H = 1$. Secondly, even if $H = 1$ is the only possible explanation of $T = 1$, this does not change the intuition that it has a very low explanatory power.

8.6 A New Proposal for Measuring Causal Explanatory Power

If none of the measures discussed in sections 8.4 and 8.5 is able to measure the power of an explanation H to explain a given explanandum E, how else should we measure it? In this section, I aim to answer this question, at least, when it comes to causal explanations. Notice that Gärdenfors (1988), McGrew (2003), Schupbach and Sprenger (2011), Crupi and Tentori (2012) all advance their measures as universal measures of explanatory power, that are supposed to be applicable to any kind of explanation, causal or non-causal. The approach, that I will propose here is somewhat more modest. It aims at determining the explanatory power of causal explanations only and it will completely abstain from making any claims about the power of non-causal explanations or whether non-causal explanations even have a similar dimension of explanatory goodness. Since the explanatory power of a causal explanation will be completely determined by the intrinsic causal properties of the events that constitute the explanation, I will, just like Eva and Stern (2019), call the explanatory power of a causal explanation its *causal explanatory power*.

I have already pointed out that, what we intuitively perceive as an explanation's power to explain a given explanandum, is independent from the degrees of belief that we ascribe to the explanation itself, to the explanandum, and to alternative explanations of the explanandum. I have argued that the power of an explanation H to explain a given explanandum E is therefore not the actual reduction of surprise about E that learning or assuming H induces in the given situation. It is rather an ability that is intrinsic to the explanation itself. The basic idea of my proposal is therefore this: For an extensive causal explanation of a given explanandum ϕ, the intrinsic ability to reduce the surprise about ϕ stems from the intrinsic causal powers of the causes that constitute the explanation.

As argued in chapter 7, any extensive causal explanation conforms to the following schema: 'ϕ, *because* $\vec{X} = \vec{x}$, *despite* $\vec{Y} = \vec{y}$'. We have seen that, which causal elements the *because*- and the *despite*-part are composed of, depends on whether the explanandum is a deviant or a default state. How exactly the causal explanatory power of an extensive causal explanation is determined therefore also depends on whether the explanandum is a deviant or a default state. As shown in chapter 7, if ϕ is a deviant state, any extensive causal explanation of ϕ can be subdivided into the following parts:

$$\phi \textit{ because } \vec{C} = \vec{c} \wedge \vec{O} = \vec{o}, \textit{ despite } \vec{P} = \vec{p} \tag{8.18}$$

where $\vec{C} = \vec{c}$ is a potential causal producer of ϕ, $\vec{O} = \vec{o}$ is a possibly empty conjunction of omissions (or preventers) of potential ($\vec{C} = \vec{c}$)-specific preventers of ϕ and $\vec{P} = \vec{p}$ is a possibly empty conjunction of potential ($\vec{C} = \vec{c}$)-specific preventers of ϕ. With this in mind, we can put forward the following definition:

Causal Explanatory Power (Deviant States). *If 'ϕ because $\vec{C} = \vec{c} \wedge \vec{O} = \vec{o}$, despite $\vec{P} = \vec{p}$' is an extensive causal explanation of the deviant state ϕ in a probabilistic SEM $(\mathcal{M}, \mathcal{P})$, where $\vec{C} = \vec{c}$ is a potential causal producer of ϕ, $\vec{O} = \vec{o}$ is a possibly empty conjunction of omissions (or preventers) of potential ($\vec{C} = \vec{c}$)-specific preventers of ϕ and $\vec{P} = \vec{p}$ is a possibly empty conjunction of potential ($\vec{C} = \vec{c}$)-specific preventers of ϕ, then its causal explanatory power is* *given by:*

$$ECI^{\vec{P}=\vec{p}}_{\vec{C}=\vec{c},\phi} \tag{8.19}$$

which is the effective causal influence of $\vec{C}=\vec{c}$ on ϕ in the presence of $\vec{P}=\vec{p}$.

As also shown in chapter 7, if ϕ is a default state, any extensive causal explanation of ϕ can be subdivided into the following elements, each of which can be empty:

$$\phi \textit{ because } \vec{O}=\vec{o} \wedge \vec{P}_1=\vec{p}_1 \wedge \vec{P}_m=\vec{p}_m, \textit{ despite } \vec{C}_1=\vec{c}_1 \wedge \vec{R}_1=\vec{r}_1 ... \wedge \vec{C}_n=\vec{c}_n \wedge \vec{R}_n=\vec{r}_n \tag{8.20}$$

$\vec{O}=\vec{o}$ is a conjunction of omissions of potential producers of a complement of ϕ, and for all i with $1 \leq i \leq m$, $\vec{P}_i=\vec{p}_i$ is a $(\vec{D}_i=\vec{d}_i)$-specific potential preventer of a complement of ϕ, with $\vec{D}_i=\vec{d}_i$ being a potential producer of a complement of ϕ that is not in $\vec{O}=\vec{o}$, and for all j with $1 \leq j \leq n$, $\vec{C}_j=\vec{c}_j$ is a potential producer of a complement of ϕ with $\vec{R}_j=\vec{r}_j$ being a conjunction of omissions (or preventers) of $(\vec{C}_j=\vec{c}_j)$-specific potential preventers of the complement of ϕ. This enables us to define:

Causal Explanatory Power (Default States). *If 'ϕ because $\vec{O}=\vec{o} \wedge \vec{P}_1=\vec{p}_1 \wedge \vec{P}_m=\vec{p}_m$, despite $\vec{C}_1=\vec{c}_1 \wedge \vec{R}_1=\vec{r}_1 ... \wedge \vec{C}_n=\vec{c}_n \wedge \vec{R}_n=\vec{r}_n$' is an extensive causal explanation of the default state ϕ in a probabilistic causal model $(\mathcal{M},\mathcal{P})$, where $\vec{O}=\vec{o}$ is a conjunction of omissions of potential producers of a complement of ϕ, and for all i with $1 \leq i \leq m$, $\vec{P}_i=\vec{p}_i$ is a $(\vec{D}_i=\vec{d}_i)$-specific potential preventer of a complement of ϕ, with $\vec{D}_i=\vec{d}_i$ being a potential producer of a complement of ϕ that is not in $\vec{O}=\vec{o}$, and for all j with $1 \leq j \leq n$, $\vec{C}_j=\vec{c}_j$ is a potential producer of a complement of ϕ with $\vec{R}_j=\vec{r}_j$ being a conjunction of omissions (or preventers) of $(\vec{C}_j=\vec{c}_j)$-specific potential preventers of the complement of ϕ, then its causal explanatory power is given by:*

$$\prod_{i\in\{1,...,m\}}(1-ECI^{\vec{P}_i=\vec{p}_i}_{\vec{D}_i=\vec{d}_i,\neg\phi}) \times \prod_{j\in\{1,...,n\}}(1-ECI^{\varnothing}_{\vec{C}_j=\vec{c}_j,\neg\phi}) \tag{8.21}$$

To get a better feel for the rather abstract definitions, let us look at a few examples. Consider first the probabilistic poisoning scenario from figure 7.2, where we have two potential assassins that may try to kill Pete. The first, assassin A, considers to poison Pete ($A=1$), while the second, assassin R, considers to shoot him ($R=1$). We assumed that ingesting the poison has, in the absence of any preventive causes, a generative causal power of 0.9 on killing the person who ingests it. We further assumed that a shot from the second assassin has a generative causal power of 0.7 on killing his target. Furthermore, Pete's bodyguard may put an antidote into Pete's tea ($B=1$) which, if administered, is able to prevent the poison from killing Pete with a preventive power of 0.8. We considered a situation, in which we have already learned that Pete died ($D=1$). We identified the following extensive causal explanations of Pete's death:

(1) $D=1$ *because* $A=1 \wedge B=0$.

(2) $D=1$ *because* $R=1$.

(3) $D = 1$ *because* $A = 1$, *despite* $B = 1$.

According to our new explication of causal explanatory power, we get the following results:

(1) '$D = 1$ *because* $A = 1 \wedge B = 0$' (Pete died because assassin A has put poison into his tea and his bodyguard did not put any antidote into his tea.) has an explanatory power of $ECI^{\varnothing}_{A=1,D=1} = 0.9$.

(2) '$D = 1$ *because* $R = 1$' (Pete died because assassin B shot him.) has an explanatory power of $ECI^{\varnothing}_{R=1,D=1} = 0.7$.

(3) '$D = 1$ *because* $A = 1$, *despite* $B = 1$" (Pete died because assassin A has put poison into his tea, despite the fact that his bodyguard has put an antidote into his tea.) has an explanatory power of $ECI^{B=1}_{A=1,D=1} = 0.9 \times (1 - 0.8) = 0.18$.

Next, consider the scenario from figure 7.3, where we assumed that there is only one potential assassin of Pete, namely assassin A, who considers to poison his tea with the poison ($A = 1$) that has a generative causal power of 0.9 on killing the person who ingests it. We additionally assumed that there are two potential preventers of Pete's assassination. First, Pete's bodyguard may put an antidote into Pete's tea ($B = 1$) which, if administered, is able to prevent the poison from killing Pete with a preventive power of 0.8. Additionally, Pete's cook may put an antidote into his tea ($C = 1$) which, if administered, is able to prevent the poison from killing Pete with a preventive power of 0.5. We considered a situation, in which we have already learned that Pete died ($D = 1$) and we identified the following extensive causal explanations of Pete's death:

(1) $D = 1$ *because* $A = 1 \wedge B = 0 \wedge C = 0$.

(2) $D = 1$ *because* $A = 1 \wedge B = 0$, *despite* $C = 1$.

(3) $D = 1$ *because* $A = 1 \wedge C = 0$, *despite* $B = 1$.

(4) $D = 1$ *because* $A = 1$, *despite* $B = 1 \wedge C = 1$.

According to our new explication of causal explanatory power, we get the following results:

(1) '$D = 1$ *because* $A = 1 \wedge B = 0 \wedge C = 0$' (Pete died because assassin A has put poison into his tea and neither Pete's bodyguard nor Pete's cook did put any antidote into Pete's tea) has an explanatory power of $ECI^{\varnothing}_{A=1,D=1} = 0.9$.

(2) '$D = 1$ *because* $A = 1 \wedge B = 0$, *despite* $C = 1$' (Pete died because assassin A has put poison into his tea and his bodyguard did not put any antidote into his tea, despite the fact that his cook has put an antidote into his tea.) has an explanatory power of $ECI^{C=1}_{A=1,D=1} = 0.9 \times (1 - 0.5) = 0.45$.

(3) '$D = 1$ *because* $A = 1 \wedge C = 0$, *despite* $B = 1$' (Pete died, because assassin A has put poison into his tea and his cook did not put any antidote into his tea, despite the fact that his bodyguard has put an antidote into his tea.) has an explanatory power of $ECI^{C=1}_{A=1,D=1} = 0.9 \times (1 - 0.8) = 0.18$.

(4) '$D = 1$ *because* $A = 1$, *despite* $B = 1 \wedge C = 1$' (Pete died, because assassin A has put poison into his tea, despite the fact that his bodyguard and his cook did put their antidotes into his tea.) has an explanatory power of $ECI^{C=1}_{A=1,D=1} = 0.9 \times (1-0.5) \times (1-0.8) = 0.09$.

Consider now the scenario from figure 7.4, which has the same causal structure as the scenario from figure 7.2, only this time we assumed that Pete continues to live ($D = 0$). We identified the following extensive causal explanations of Pete's survival:

(1) $D = 0$ *because* $A = 0 \wedge R = 0$.

(2) $D = 0$ *because* $B = 1 \wedge R = 0$.

(3) $D = 0$ *because* $B = 1$, *despite* $R = 1$.

(4) $D = 0$ *because* $A = 0$, *despite* $R = 1$.

(5) $D = 0$ *because* $R = 0$, *despite* $A = 1 \wedge B = 0$.

(6) $D = 0$ *despite* $A = 1 \wedge B = 0 \wedge R = 1$.

According to our new explication of causal explanatory power, we get the following results:

(1) '$D = 0$ *because* $A = 0 \wedge R = 0$' (Pete lives, because assassin A has not put poison into his tea and assassin B has not shot on him.) has an explanatory power of 1.

(2) '$D = 0$ *because* $B = 1 \wedge R = 0$' (Pete lives, because his bodyguard has put an antidote into his tea and assassin B has not shot on him.) has an explanatory power of $(1 - ECI^{B=1}_{A=1,D=1}) = 1 - 0.9 \times (1 - 0.8) = 0.82$

(3) '$D = 0$ *because* $B = 1$, *despite* $R = 1$' (Pete lives, because his bodyguard has put an antidote into his tea, despite the fact that assassin B has shot on him.) has an explanatory power of $(1 - ECI^{B=1}_{A=1,D=1}) \times (1 - ECI^{\varnothing}_{R=1,D=1}) = (1 - 0.9 \times (1 - 0.8)) \times (1 - 0.7) = 0.246$.

(4) '$D = 0$ *because* $A = 0$, *despite* $R = 1$' (Pete lives, because assassin A has not put poison into his tea, despite the fact that assassin B has shot on him.) has an explanatory power of $(1 - ECI^{\varnothing}_{R=1,D=1}) = 1 - 0.7 = 0.3$.

(5) '$D = 0$ *because* $R = 0$, *despite* $A = 1 \wedge B = 0$' (Pete lives, because assassin B has not shot on him, despite the fact that assassin A put poison into his tea and his bodyguard has not put an antidote into his tea.) has an explanatory power of $(1 - ECI^{\varnothing}_{A=1,D=1}) = 1 - 0.9 = 0.1$.

(6) '$D = 0$ *despite* $A = 1 \wedge B = 0 \wedge R = 1$' (Pete lives, despite the fact that assassin A put poison into his tea, his bodyguard has not put an antidote into his tea, and assassin B has shot on him.) has an explanatory power of $(1 - ECI^{\varnothing}_{A=1,D=1}) \times (1 - ECI^{\varnothing}_{R=1,D=1}) = (1 - 0.9) \times (1 - 0.7) = 0.03$.

Notice that there is no explanatory old evidence problem for our concept of causal explanatory power, since our concept of causal explanatory power is solely based on the intrinsic causal powers of the causes involved in the explanation. In chapters 4 and 5 we have already argued that generative and preventive causal powers are independent from the prior probabilities of the respective power-bearers, from the prior probability of the effect, and from the probabilities and powers of alternative causes. Consequently, our concept of causal explanatory power is independent from the probability of the explanans, from the probability of the explanandum, and from the probabilities and powers of alternative explanations of the given explanandum. Just consider the examples in section 8.4.2, that are built on the disjunctive forest fire scenario. In all three examples, the intrinsic causal power of $A = 1$ on $F = 1$ remains the same: 0.9. According to our explication of causal explanatory power, the causal explanatory power of the extensive causal explanation '$F = 1$ *because* $A = 1$" is therefore given by $ECI^{\varnothing}_{A=1,F=1} = 0.9$ in all three examples. Confounding is also no problem for my explication of causal explanatory power, since the statistical methods for determining generative and preventive causal powers are designed to avoid confounding.

Finally, let us check how our explication of causal explanatory power deals with Cohen's coin-tossing example. A single coin toss, represented by H_1, can be seen as a potential causal producer of the event E_1 = *the coin lands on heads*. Tossing the biased coin once (H_1) has a generative causal power of 0.7^1 on E_1. Tossing the biased coin k-times (H_k) has a generative causal power of 0.7^k on E_k. This gives us the results as presented in table 8.3.

Number of flips (k)	Causal Explanatory Power of H_k on E_k
1	$0.7^1 = 0.7$
2	$0.7^2 = 0.49$
3	$0.7^3 = 0.343$
4	$0.7^4 = 0.2401$
5	$0.7^5 = 0.168$
10	$0.7^{10} = 0.028$
20	$0.7^{20} = 0.0008$
30	$0.7^{30} = 0.00002$

Table 8.3: Values of causal explanatory power in Cohen's coin-tossing example.

The results are in accord with the intuition that the explanatory power should decrease with increasing values of k and, unlike ε_{CT}, our measure of causal explanatory power does not ascribe unintuitively low values of explanatory power in cases where k is small.

Before I conclude, I would like to briefly address a potential objection. My argumentation against the Bayesian measures and in favor of my new account of explanatory power is solely based on intuitions about the concept of explanatory power, where the intuitions are either my own or of a selected group of philosophers, like Glymour (2015) or Lipton (2004). But I have not based my arguments on any empirical studies that examine how human beings in general understand and assess the concept of explanatory power. This is for a good reason. While I definitely agree, that empirical studies about how humans actually understand and assess explanatory power would be highly illuminating, a lot of work still needs to be done in order to get some reliable results. Schupbach (2011a), for example, conducts an empirical study, whose results seem to support the superiority of ε_{SS} as a measure of explanatory power when compared to some alternative Bayesian measures. But (Glymour, 2015) puts forward a convincing criticism of the study and Schupbach's interpretation of the results. A central point made by Glymour is that the questions posed to the participants in the study did not ensure whether the participants were really evaluating the explanatory power of the presented explanations or the posterior probability of the explanandum under the assumption of the explanations. The empirical studies carried out by Colombo et al. (2016) seem to speak against the Bayesian measures and in favor of my new proposal for measuring causal explanatory power. They record, for example, that the prior probabilities of the presented explanation candidates, which were given in terms of objective base rates, did not influence the explanatory power assessments of the participants.[30] But here again the studies are problematic. The vignettes that were used in the study, could very well be interpreted as asking about ones certainty about whether a given hypothesis is an explanation, instead of asking about the explanatory power of the hypothesis.[31] Thus, as long as reliable results from empirical studies are still lacking, thoughtful individual intuitions about explanatory power are the best guideline for building a formal account of explanatory power.

8.7 Summary

The present chapter has been concerned with evaluating causal explanations. Just like Schupbach and Sprenger (2011), I followed the intuition that an explanation's "ability to decrease the degree to which we find the explanandum surprising" (Schupbach and Sprenger, 2011, p. 108) amounts to an important dimension of explanatory goodness. I have shown that the three Bayesian measures ε_{SS}, ε_{CT}, and ε_{GMG} face significant problems, which strongly suggest that none of the three measures is able to capture, what we intuitively understand under the concept of explanatory power. I have shortly addressed some further measures of explanatory power, that have been proposed in the philosophical literature, and I have argued that they face very similar problems. Finally, I have put forward a new approach to measure the explanatory power of causal explanations. This approach assumes that the explanatory power of a causal explanation emerges from the intrinsic causal powers of the causes that compose the explanation.

[30] See (Colombo et al., 2016, p. 436).

[31] See (Colombo et al., 2016, p. 434). For recording assessments of explanatory power of a given hypothesis on a given explanandum, the participants were asked to assess, on a scale from 1 to 7, how much they agree whether the hypothesis explains the explanandum.

Part III

Proportional Causal Explanations

Chapter 9

Causation on Several Levels

9.1 Introduction

I have based my account of causal explanation on the framework of causal models, because the framework has come a very long way in representing causal relationships in a formally rigorous way. It enables to formulate precise definitions of causal concepts and to evaluate causal claims unambiguously relative to a given causal model. However, the framework has a deficiency that I have ignored so far. The formalism of causal models and the interventionist accounts of causation that are built on it, have long neglected non-causal relationships of counterfactual dependence, like supervenience, grounding, realization, or determination. This neglect results in the deficiency that intuitively adequate causal claims on different supervenience levels about a single causal scenario cannot be evaluated in a single causal model. I will call this deficiency the supervenience related expressivity problem of causal models, or in short: *the expressivity problem.*

In this chapter, I will propose a new solution to the expressivity problem. But before I can do so, I have to address the question of whether the expressivity problem even is a problem in the first place. Causal exclusion arguments contend that the restriction of causal claims on only one supervenience level is a virtue and not a vice. But even though they sound convincing at first, I will argue that causal exclusion arguments do not apply to interventionist concepts of causation. After having established that the expressivity problem really is a problem for causal-model-based accounts of causation, I will discuss two already existing proposals to solve the problem. I will argue that both proposals have significant downsides. I will therefore propose a new solution that consists in an extension of the classical causal model framework.

9.2 The Expressivity Problem

The expressivity problem is best illustrated with an example. Consider the pigeon Sophie, who rose to stardom after having been introduced in an example by Yablo (1992). Imagine that Sophie lives in a laboratory where she was trained to peck if and only if a red object is shown to her. We are now interested in a specific situation, in which Sophie may be confronted with objects of a wide variety of very specific colour shades. For example, among the red objects are objects that are crimson, scarlet, maroon, or rose. Among the blue objects, there are objects that are indigo, azure, navy, or cyan, and so on. The causal relationship between Sophie's pecking behaviour and the colour of the object that is shown to her can be represented by a causal model $\mathcal{M}^S$ that contains two variables. Variable C represents the colour of the object shown to Sophie and variable P represents whether Sophie pecks. The value set of C contains all the available colour shades: $\mathcal{R}(C)$ = {crimson, scarlet, maroon, rose, indigo, azure, navy, cyan, ...}. Variable P, on the other hand, has only two values: $P = 1$ represents that Sophie pecks and $P = 0$ represents that Sophie does not peck.

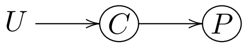


Figure 9.1: Causal Diagram for the Sophie scenario.

The structural equation that describes the causal relationship between C and P is given by:

- $P = 1$ iff $C = \textit{crimson} \lor C = \textit{scarlet} \lor C = \textit{maroon} \lor C = \textit{rose}$

Now, imagine a token situation $(\mathcal{M}^S, u)$, in which the object shown to Sophie is scarlet, that is $(\mathcal{M}^S, u) \models C = \textit{scarlet}$. Well-trained Sophie will accordingly peck, that is $(\mathcal{M}^S, u) \models P = 1$. By using the HP-definition of actual causation, we can evaluate the following causal claim in $(\mathcal{M}^S, u)$, which turns out to be true: '$C = \textit{scarlet}$ (the object being scarlet) is an actual cause of $P = 1$ (Sophie's pecking)'. But this is not the only intuitively adequate causal claim about the given situation. Here is another one: 'the object being red is an actual cause of Sophie's pecking'. The claim is intuitively adequate because the scarlet object is indeed red, Sophie does indeed peck, and there is a counterfactual de facto dependence between Sophie's pecking and the object being red that, according to the HP-definition, amounts to a relation of actual causation. But here is the problem: The claim 'the object being red is an actual cause of Sophie's pecking' cannot be evaluated in the given causal model for the simple reason that 'the object is red' is not explicitly represented as an event in $\mathcal{M}^S$.

Before discussing this problem further, it is useful to first explain the relationship between red and scarlet in a bit more detail. Scarlet and red are typical examples of a relation that is commonly known as *determination*: The property *scarlet* is a determinate of the determinable *red*. Determination is a non-causal relation of counterfactual dependence that Yablo describes like this: "P determines Q iff to possess the one is to possess the other, not simpliciter, but in a certain way" (Yablo, 1992, p. 260). Although determination is often considered to apply to properties, it can easily be applied to events as well. Yablo reformulates his description accordingly: "p determines q iff: for p to occur (in a possible world) is for q to occur (there), not simpliciter, but in a certain way"(Yablo, 1992, p. 260). Even though I will focus on examples that contain determination relations, the expressivity problem and the solution, that I will later propose, apply to other kinds of non-causal relations of counterfactual dependence, like supervenience, grounding, or realization as well. A detailed characterization or differentiation of all these relations will not be necessary for the purposes of the present chapter. In the following, whenever I speak of determination or supervenience, I always mean a relation of non-causal, counterfactual dependence that satisfies at least the following conditions:[1]

[1] Wilson (2021) formulates a list of features that are typically ascribed to the relation of determination. A

- *Transitivity*: If C is a determinable of B and B is a determinable of A, then C is a determinable of A.
- *Relativity*: An event A can be a determinable of one event B while being a determinate of another event C.
- *Upwards Enheritance*: For every determinable A of a determinate B: if B occurs at a time t and place s then A occurs at t and s.
- *Unique Determinable under Alternatives*: If an event A is a determinable of B, then no alternative to A (an event that is incompatible with A) can also be a determinable of B. So, according to *Upwards Enheritance*, there cannot be a change of A without a change of B.

It is this last feature that suggests the common talk of supervenience- or determination-levels. If A is a determinable of B, then from B's point of view, A and all its alternatives form a level. If there are two events that are both determinables of B, then they must lie on different levels. A, on the other hand, can very well be a determinable of several alternatives to B. This is why a change in B does not necessarily imply a change in A. Every level is therefore a set of mutually exclusive events, such that each event on a given level has only one determinable/supervening event at every higher level, but potentially several determinates/supervenience bases at every lower level. The last feature, which I assume to hold for the supervenience relations that I aim to consider here, incorporates the idea of levels:

- *Requisite and Unique Determination*: If A occurs at a time t and place s, then for every level L of determinates of A: one and only one L-level determinate B of A occurs at t and s.

Having clarified the relation between scarlet and red, we can now reformulate the expressivity problem in the Sophie-scenario in the following way: In the situation under consideration, there are at least two causal claims on different determination levels that appear to be intuitively adequate. But in the given causal model our expressivity is restricted to one determination level only. One might think that the problem can be fixed very easily. If the causal claim 'the object being red is an actual cause of Sophie's pecking' cannot be evaluated in the given causal model for the simple reason that 'the object is red' is not explicitly represented as an event in $\mathcal{M}^S$, why then don't we just adapt $\mathcal{M}^S$ to explicitly include 'the object is red' as an event in the model? The reason is that there are two well-entrenched rules of causal modeling that speak against it.

Probably the easiest way to integrate the event 'the object is red' into $\mathcal{M}^S$, would be to just add *red* to the value set $\mathcal{R}(C)$ of C. But here is the causal modeling rule, that speaks against it:

Mutual Exclusivity. *For every variable V in a causal model $\mathcal{M}$: Any two values of V are incompatible with each other.*

comparison shows that the first four conditions that I assume to hold for the supervenience relations under consideration amounts to a small subset of these features. Wilson does not explicitly mention the last feature (*Unique Determinable under Alternatives*), though. But it clearly holds for determination.

Like most rules of causal modeling, Mutual Exclusivity is rarely explicitly discussed, but it is actually always tacitly assumed in the causal model literature. And it clearly makes sense to do so. If we cannot presuppose Mutual Exclusivity for a causal model $\mathcal{M}$, then the information that a variable V in $\mathcal{M}$ takes on a certain value does not exclude the possibility that V additionally takes on some other values. A complete description of a causal setting would therefore become much more complex, since we would need to state for every value v in $\mathcal{R}(V)$, whether $V = v$ holds or not. Now, adding *red* to $\mathcal{R}(C)$ would clearly violate Mutual Exclusivity, because C would have several pairs of values, like *red* and *scarlet*, that are not incompatible with each other.

Here is another way to integrate the event 'the object is red' into $\mathcal{M}^S$. In addition to variable C, we could introduce another variable $\tilde{C}$ with the value set $\mathcal{R}(\tilde{C}) = \{red, \neg red\}$ into $\mathcal{M}^S$ and the following structural equation, which describes the causal relationship between $\tilde{C}$ and P:

- $P = 1$ iff $\tilde{C} = red$

The causal diagram of this extended model, which I will denote by $\mathcal{M}^{\tilde{S}}$, is presented in figure 9.2.

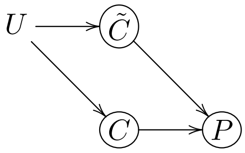


Figure 9.2: Causal diagram for $\mathcal{M}^{\tilde{S}}$.

The events 'the object is red' ($\tilde{C} = red$) and 'the object is *scarlet* ($C = scarlet$) are now both explicitly represented in $\mathcal{M}^{\tilde{S}}$, without a violation of Mutual Exclusivity. But yet again, we violate a commonly presupposed rule of causal modeling that Woodward (2015) has dubbed *Independent Fixability* and that he formulates as follows:[2]

Independent Fixability (IF). *A causal model $\mathcal{M}$ with variable-set $\mathcal{V}$ satisfies Independent Fixability (IF) if and only if for each variable $V \in \mathcal{V}$ and each $v \in \mathcal{R}(V)$ it is possible (in terms of their assumed definitional, logical, mathematical, mereological or supervenience relations) to set V on v via an intervention, concurrently with each of the other variables in $\mathcal{V}$ also being set to any of its individually possible values by independent interventions.*

In $\mathcal{M}^{\tilde{S}}$, $\tilde{C}$ and C are not independently fixable, because they represent events that are related in terms of determination. It is therefore metaphysically impossible to change the value of $\tilde{C}$ without also changing the value of C.

So, as soon as we explicitly incorporate the event 'the object is red' into the causal model $\mathcal{M}^S$, we violate some established rule of causal modeling. This brings us to an obvious question: Why then don't we just break an established rule of causal modeling? If we want to be able to express and to evaluate causal claims on several supervenience levels in one and the same causal model, this is indeed what we will have to do. But breaking well-established rules of

[2] See (Woodward, 2015, p. 316).

causal modeling will have consequences and what those consequences are, clearly depends on which rule we will break. The essential task in solving the expressivity problem will therefore be to explore, which rule violation brings the most benefits and the fewest problems.[3] But before I tackle this task, I should first deal with an argument, according to which the expressivity problem does not even need a solution, since its existence is more of a virtue than a vice.

9.3 The Causal Exclusion Argument

The causal exclusion argument has been put forward by Jaegwon Kim to contest the existence of mental causation, given the assumption of non-reductive physicalism.[4] But the argument is not restricted to the supervenience of the mental on the physical. It can be applied to any kind of supervenience relation that satsifies the features listed above and the additional charactersitic of non-reducibility:

- *Non-Reducibility*: A supervening event is metaphysically distinct from (non-reducible to) any of its supervenience bases.

I do not want to resolve the question of whether or in which cases non-reducibility actually holds for a supervenience relation. My aim is rather to argue that, even if non-reducibility does hold for a supervenience relation under consideration, the causal exclusion argument does still not apply to interventionist concepts of causation. In the following discussion, I will therefore simply assume non-reducibility, just for the sake of the argument.

Here is how the argument goes: Imagine that event $C = c$ is causally sufficient for another event $E = e$. Additionally, an event $\widetilde{C} = \tilde{c}$, which supervenes on $C = c$, causes an event $\widetilde{E} = \tilde{e}$, which supervenes on $E = e$. The scenario can be illustrated by the diagram in figure 9.3.

$$\begin{tikzcd}
\boxed{\widetilde{C} = \tilde{c}} \arrow[r] & \boxed{\widetilde{E} = \tilde{e}} \\
\boxed{C = c} \arrow[r] \arrow[u, Rightarrow] & \boxed{E = e} \arrow[u, Rightarrow]
\end{tikzcd}$$

Figure 9.3: Scenario used for the causal exclusion argument. Adapted from (Kim, 2005, p. 63).

Since $\widetilde{E} = \tilde{e}$ supervenes on $E = e$, $E = e$ necessitates $\widetilde{E} = \tilde{e}$ according to *Upwards Inheritance*. Since $C = c$ is causally sufficient for $E = e$, $C = c$ is consequently also causally sufficient for

[3] One might hope to solve the expressivity problem by simply identifying the event 'the object is red' with the disjunction of all the events, in which the object takes on some specific shade of red. Halpern and Pearl's (2005a) definition of actual causation does not allow disjunctive causes, though. Halpern and Pearl (2005a) do not even allow disjunctions in the antecedents of their intervention counterfactuals. Briggs (2012) generalizes Halpern and Pearl's semantics for intervention counterfactuals to also allow disjunctions in the antecedent. This can be seen as a possible solution to the expressivity problem, as Briggs explicitly points out (Briggs, 2012, cf. p. 149). But it is a problematic one for several reasons. First, it is a controversial claim that a determinable or a supervening event is identical to the disjunction of its determinates or supervenience bases. Secondly, even if such an identity holds, the knowledge of this identity is simply presupposed in Briggs' approach and not explicitly represented in the causal model. So, as long as these identities are not explicitly incorporated into the model, only the causal claim 'the object being scarlet or crimson or maroon or rose is an actual cause of Sophie's pecking' can be evaluated and expressed in the causal model, but not the claim 'the object being red is an actual cause of Sophie's pecking'.

[4] See, for example, (Kim, 2005).

$\widetilde{E} = \tilde{e}$. Since $\widetilde{C} = \tilde{c}$ is metaphysically distinct from $C = c$, it follows from the assumption that $\widetilde{C} = \tilde{c}$ also causes $\widetilde{E} = \tilde{e}$, that $\widetilde{E} = \tilde{e}$ is causally overdetermined by the two events $C = c$ and $\widetilde{C} = \tilde{c}$. Since the presented causal schema is a very common scenario, either causal overdetermination is very common or the supposition that $\widetilde{C} = \tilde{c}$ is a cause of $\widetilde{E} = \tilde{e}$ is false.[5]

The argument works analogously for a scenario as shown in figure 9.4, where $\widetilde{C} = \tilde{c}$ is supposed to cause $E = e$, the effect of $C = c$, itself, instead of an event $\widetilde{E} = \tilde{e}$ that supervenes on $E = e$.

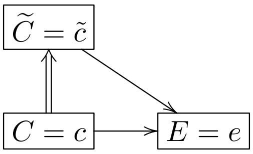


Figure 9.4: Scenario of downwards causation. Adapted from (Baumgartner, 2013, p. 7).

In this case, the trouble comes in form of an unintuitive overdetermination of $E = e$ by the metaphyically related but distinct events $\widetilde{C} = \tilde{c}$ and $C = c$. Sophie can once again serve as an example. The object being scarlet is causally sufficent for Sophie to peck. The object being red supervenes on the object being scarlet. But despite their metaphysical relation due to supervenience, both events are metaphysically distinct. Now, suppose that the object being red also causes Sophie to peck, then Sophie's pecking is causally overdetermined, since it is caused by two metaphysically distinct events: The object being scarlet and the object being red. It seems that we are facing the following decision: Either we accept, contrary to intuition, that Sophie's pecking is causally overdetermined, or we accept that only one of the two causal claims is true and that causation only emanates from one of the two supervenience levels. Causal exclusion is the position that opts for the latter, which is why, according to this position, the expressivity problem of causal models is not a problem that should be eliminated, but a virtue that should be preserved.

I do not want to claim that the position of causal exclusion is generally inadequate. Instead, I agree with Hitchcock (2012), that whether the position of causal exclusion is adequate or not, highly depends on the concept of causation under consideration. And when it comes to interventionist concepts of causation, a crucial premise of the causal exclusion argument goes amiss. To see this, let us first list all the key premises of the causal exclusion argument:

Non-Reducibility: The supervening event $\widetilde{C} = \tilde{c}$ is metaphysically distinct from its supervenience base $C = c$.

Upwards Enheritance: When $\widetilde{C} = \tilde{c}$ supervenes on $C = c$, then the occurrence of $C = c$ necessitates the occurrence of $\widetilde{C} = \tilde{c}$.

Causal Sufficiency: When an event $C = c$ is causally sufficient for an event $E = e$, then $C = c$ is also causally sufficient for the event $\widetilde{E} = \tilde{e}$ that supervenes on $E = e$.

[5] There are several slightly different formulations of the causal exclusion argument in the literature. See, for example, (Kim, 2005), (Woodward, 2015), (Weslake, 2011), (Hitchcock, 2012), (Gebharter, 2017), (Polger et al., 2018). My formulation may differ from other formulations in some respects. But the crucial point is always that the assumption of causation on several supervenience levels would lead to an unintuitive instance of overdetermination.

Overdetermination: When an event $C = c$ is causally sufficient for an event $\widetilde{E} = \tilde{e}$ and another event $\widetilde{C} = \tilde{c}$, that is metaphysically distinct from $C = c$, causes $\widetilde{E} = \tilde{e}$, then $\widetilde{E} = \tilde{e}$ is causally overdetermined.

Rarity: Causal Overdetermination is rare.

We have already assumed that *Upwards Enheritance* holds for the supervenience relations that we consider here and I wanted to presuppose *Non-Reducibility* for the sake of the argument. I have also no quarrel with *Causal Sufficiency* and *Rarity.* Both seem to be rather fair assumptions. The premise that I have a quarrel with is *Overdetermination. Overdetermination* is completely unfounded when it comes to mere dependence-concepts of causation, like the ones developed by interventionist accounts of causation. If we employ a dependence-concept of causation, like the HP-definition of actual causation, the claim that $\widetilde{C} = \tilde{c}$ is a cause of $\widetilde{E} = \tilde{e}$ only means that the occurrence of $\widetilde{E} = \tilde{e}$ counterfactually de facto depends on the occurrence of $\widetilde{C} = \tilde{c}$. The causal claim does not entail the assertion that there is a separate causal process, an energy-transference or a mechanism emanating specifically from $\widetilde{C} = \tilde{c}$ to $\widetilde{E} = \tilde{e}$. We can call such a causal claim, namely one that only asserts a certain counterfactual dependence between cause and effect, a *thin causal claim*, while a causal claim that asserts that there is a separate causal process, energy-transference or mechanism emanating specifically from the cause-event to the effect-event, is a *thick causal claim.*[6]

Now, as Woodward (2015) points out, it is the presence of two separate causal processes or mechanisms yielding the same effect, which constitutes causal overdetermination.[7] This is why the premise *Overdetermination* only makes sense, if we understand both causal claims, namely that $C = c$ is causally sufficient for $\widetilde{E} = \tilde{e}$, while $\widetilde{C} = \tilde{c}$ also causes $\widetilde{E} = \tilde{e}$, as thick causal claims. But if both causal claims are thin causal claims, which is the case, if we employ an interventionist concept of causation, then both causal claims can be true, even if there is only one separate causal process or mechanism that yields $\widetilde{E} = \tilde{e}$. *Overdetermination* does not hold in such a case.[8] Coming back to Sophie: Using an interventionist concept of causation, we can claim both that the object being scarlet causes Sophie to peck and that the object being red causes Sophie to peck, without thereby claiming that Sophie's pecking is causally overdetermined. The two claims only assert that there is a counterfactual dependence between Sophie's pecking and the object being scarlet and a counterfactual dependence between Sophie's pecking and the object being red. Both counterfactual dependencies may very well be grounded in one and the same causal process or mechanism.

Since the causal exclusion argument does not give us any reason to ban thin causal claims on several supervenience levels, the causal exclusion argument does not give us any reason to think that the expressivity problem is a virtue, rather than a vice. It is therefore now time to seek a solution for it.

[6] This terminology is adopted from Woodward's (2015) differentiation between thin and thick conceptions of causation. Notice though, that this differentiation does not necessarily coincide with the distinction between mere dependence-causation and causal production. In chapter 6, I have argued that causal production is not necessarily constituted by a separate process, energy-transference or mechanism.

[7] See (Woodward, 2015, p. 344).

[8] The same argument is put forward by (Woodward, 2015) and by (Polger et al., 2018, cf. p. 54).

9.4 Proposals to Address the Expressivity Problem

9.4.1 Dropping Independent Fixability

As already pointed out in section 9.2, a natural way to take care of the expressivity problem is to drop the condition of Independent Fixability (IF) and to allow causal models to contain separate variables that are related in terms of supervenience. In the Sophie-scenario, $\mathcal{M}^{\tilde{S}}$ could therefore be acknowledged as a permissible causal model and we would be able to explicitly represent events that are related in terms of supervenience, like 'the object is scarlet' and 'the object is red', in a single causal model. This is essentially how Woodward (2015) addresses the expressivity problem. But in dropping IF some difficulties arise that we have to consider.

Since IF has been a default assumption for most interventionist accounts of causation, standard definitions of causation within the interventionist paradigm turn out to be inadequate in models that do not fulfill IF. For example, with the presupposition of IF, no interventionist definition of causation needed to be sensitive enough to distinguish a supervenience relation from a causal relation. Consider, for example, a setting $(\mathcal{M}^{\tilde{S}}, u)$ with $C = \textit{scarlet}$ and $\tilde{C} = \textit{red}$. Due to the counterfactual dependence between these two events, the HP-definition of actual causation in its current form identifies $C = \textit{scarlet}$ as an actual cause of $\tilde{C} = \textit{red}$ and $\tilde{C} = \textit{red}$ as an actual cause of $C = \textit{scarlet}$. And there is another problem. Interventionist concepts of causation demand that for $\tilde{C} = \tilde{c}$ to be identified as a cause of $E = e$, there must be a theoretically possible intervention on $\tilde{C}$, which changes the value of $\tilde{C}$ and thereby also induces, at least in a certain contingency, a change in the value of E. But according to the standard constraints on interventions,[9] the intervention on $\tilde{C}$ must not influence any other variable with a causal influence on $E = e$, unless this influence is causally mediated by the value change of $\tilde{C}$. But, as Baumgartner (2010) points out, if there is variable C in the causal model, whose values form supervenience bases of the values of $\tilde{C}$, then any intervention that changes the value of $\tilde{C}$ also changes the value of C, which is a change that is not causally mediated by the value change of $\tilde{C}$. According to the standard constraints on interventions, there is therefore no theoretically possible intervention on $\tilde{C}$, which changes the value of $\tilde{C}$ and thereby also induces, at least in a certain contingency, a change in the value of E. As a consequence, it is impossible for $\tilde{C} = \tilde{c}$ to be a cause of $E = e$, according to the standard constraints on interventions.

This all shows that dropping the condition of IF as a rule of causal modeling would force us to make significant adjustments to the standard interventionist definitions of causation, including the definition of an intervention itself. Otherwise, interventionist definitions of causation would be unable to identify causal relations in IF-violating causal models. The definitions would confuse supervenience relations with causal relations and they would be unable to recognize any events as causes, whose supervenience bases are also explicitly represented in the model.[10]

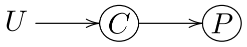
[9] See, for example, Woodward's definition of an intervention in (Woodward, 2003, p. 98).

[10] Baumgartner (2010), (2013), (2018) takes this as an argument that interventionist accounts of causation ultimately lead to the position of causal exclusion. This does of course only hold, if we would retain the standard interventionist definitions, despite dropping IF as a condition on causal modeling. Baumgartner argues that we should indeed do so, while Woodward (2015) argues for changing the standard interventionist definitions accordingly. Both are viable options. But each path leads to a different conception of causation. Baumgartner's path does indeed lead to a position of causal exclusion in an interventionist framework of causation, which is reasonable, as long as one aims to capture 'thick' causal claims. Woodward's path, on the other hand, is the right choice as long as one aims to capture 'thin' causal claims, which has traditionally been the goal of interventionist

Even though this is definitely a challenge, it is not an insurmountable one. Woodward (2015) illustrates how several well-entrenched interventionist definitions could be adjusted to perform adequately in IF-violating causal models. Following Woodward's approach would lead us to definitions of actual causation, that would adequately identify both the object being scarlet and the object being red as actual causes of Sophie's pecking in $\mathcal{M}^{\tilde{S}}$. But, nonetheless, I still consider this approach to be an inappropriate solution to the expressivity problem. The reason is, that the resulting IF-violating causal models still have no explicit representation of the supervenience relations between the events in the represented causal scenario. We are able to represent that both 'the object is scarlet' and 'the object is red' are causes of Sophie's pecking, but we are still unable to represent that 'the object is scarlet' is a determinate of 'the object is red' and that the two relations of actual causation just mentioned are therefore not entirely unrelated.

9.4.2 Biting the Bullet

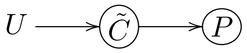
Another way to address the expressivity problem is to simply bite the bullet. This is basically the path chosen by Eronen and Brooks (2014) and Polger et al. (2018). The authors hold on to Mutual Exclusivity and IF, and they accept the fact that a causal model can only encompass one supervenience level at a time. If we want to evaluate causal claims on different supervenience levels, like 'the object being scarlet causes Sophie to peck' and 'the object being red causes Sophie to peck', then we have to do so relative to two separate causal models. One with the variable C and with $\mathcal{R}(C) = \{$*crimson, scarlet, maroon, rose, indigo, azure, navy, cyan, ...*$\}$:



- $P = 1$ iff $C = \textit{crimson} \lor C = \textit{scarlet} \lor C = \textit{maroon} \lor C = \textit{rose}.$

Figure 9.5: First causal model of the Sophie scenario

And another one with the variable $\tilde{C}$ and with $\mathcal{R}(\tilde{C}) = \{\text{red}, \neg \text{ red}\}$:



- $P = 1$ iff $C = \textit{red}.$

Figure 9.6: Second causal model of the Sophie scenario

Now, it is true that relative to a given causal model, our ability to express intuitively adequate causal claims is severely limited, since we are restricted to only one supervenience level. But one could argue that this is not so bad after all, since we have access to arbitrarily many causal models, each of which can represent a certain causal relationship on a different supervenience level. Our ability to create a causal model for any supervenience level of interest basically offsets the restricted expressivity in each causal model.

But yet again, just as with Woodward's approach to the expressivity problem, we face the disadvantage that supervenience relations between events are not explicitly modeled at all. Yet

accounts of causation.

again the approach suggests that the two causal relationships, the one between 'the object is scarlet' and 'Sophie pecks', and the one between 'the object is red' and 'Sophie pecks', are entirely unrelated, which is clearly not the case.

9.5 A New Solution to the Expressivity Problem

I will now put forward a different approach to deal with the expressivity problem, one that enables to evaluate causal claims on different supervenience levels relative to a single causal model and to explicitly represent which events in the model are related in terms of supervenience. To achieve this, I will make use of a tool kit that has been introduced by List (2019): *Systems of levels.* In its most general form a system of levels is a structure of the following kind:[11]

System of Levels (List, 2019). *A system of levels is a pair* $(\mathcal{L}, \mathcal{S})$ *with*

- $\mathcal{L}$ *being a class of objects called levels*
- $\mathcal{S}$ *being a class of mappings between levels, called supervenience mappings, where each such mapping* σ *has a source level* L *and a target level* L' *and is denoted* $\sigma : L \to L'$

and the following conditions hold:

(S1) $\mathcal{S}$ *is closed under composition of mappings, i.e., if* $\mathcal{S}$ *contains* $\sigma : L \to L'$ *and* $\sigma' : L' \to L''$, *then it also contains the composite mapping* $\sigma \cdot \sigma' : L \to L''$ *defined by first applying* σ *and then applying* σ' *(where composition is associative)*

(S2) for each level L, *there is an identity mapping* $\mathbf{1}_L : L \to L$ *in* $\mathcal{S}$ *such that, for every mapping* $\sigma : L \to L'$, *we have* $\mathbf{1}_L \cdot \sigma = \sigma = \sigma \cdot \mathbf{1}_L$

(S3) for any pair of levels L *and* L', *there is at most one mapping from* L *to* L' *in* $\mathcal{S}$

Levels, the elements of $\mathcal{L}$, are sets, whose elements are arguments or values (or both) of the supervenience mappings in $\mathcal{S}$. As the name 'supervenience mapping' suggests, the intended interpretation of a mapping $\sigma : L \to L'$ in $\mathcal{S}$ is the following: If $\sigma(a) = b$ with $a \in L$ and $b \in L'$, then b supervenes on (is a determinable of) a and a is a supervenience base (determinate) of b. I will say that a level $L \in \mathcal{L}$ is *lower*, *more specific* or *more fine grained* than a level $\tilde{L} \in \mathcal{L}$ ($L < \tilde{L}$) if and only if there is a function $\sigma \in \mathcal{S}$ with $\sigma : L \to \tilde{L}$ and σ is not an identity mapping. Because of S1 the *lower*-relation is transitive.

I will use a special kind of system of levels, which I will call a *system of value set levels* (SVSL), to extend the framework of a causal model. While the signature of a classical causal model is a triple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, consisting of exogenous variables $\mathcal{U}$, endogenous variables $\mathcal{V}$, and a function $\mathcal{R}$ that assigns to every variable in $\mathcal{U} \cup \mathcal{V}$ a value set, our newly amended causal models, which I will call *causal SVSL-models*, have signatures of the form $(\mathcal{U}, \mathcal{V}, \mathcal{R}_{Lev})$, with $\mathcal{R}_{Lev}$ being a function that assigns to every element in $\mathcal{U} \cup \mathcal{V}$ a system of value set levels that is defined in the following way:

[11] See (List, 2019, p. 854).

System of Value Set Levels (SVSL). *A system of value set levels for a variable V in a causal model $\mathcal{M}$ is a pair $(\mathcal{L}_V, \mathcal{S}_V)$ with*

- *$\mathcal{L}_V$ being a class of levels, where each level $L \in \mathcal{L}_V$ is a value set for V*
- *$\mathcal{S}_V$ being a class of supervenience mappings between levels in $\mathcal{L}_V$*

and the following conditions hold:

- *$\mathcal{S}_V$ satisfies S1, S2, S3*
- *Every $\sigma \in \mathcal{S}_V$ is surjective and*
- *There is one and only one level $L_0 \in \mathcal{L}_V$ such that for any level $L' \in \mathcal{L}_V$, there is a function $\sigma : L_0 \to L'$ in $\mathcal{S}_V$. L_0 is called the fundamental value set for V, on which all other levels for V supervene. The supervenience mapping $\sigma : L_0 \to L$ from L_0 to level $L \in \mathcal{L}_V$ is denoted by: σ_0^L*

A system of value set levels is, quite obviously, a system of levels according to List's definition. But an SVSL has to satisfy two additional constraints. First, any SVSL must have a fundamental level. This constraint does not come with any ontological claims. The fundamental value set of a variable in a causal model is only fundamental in the sense, that we, as the designers of the model, have decided that this the most specific level of representation. This does not mean that it is also the most specific or fundamental level ontologically or that there even is any fundamental ontological level at all. The other additional constraint on an SVSL is that all its supervenience mappings must be surjective. This constraint ensures that every higher-level value of a variable does have at least one supervenience base (determinate) on every lower level. Since every supervenience mapping is also a function, it follows that every level $\tilde{L}$ that is higher than another level L in an SVSL induces a partition of L. To see this, imagine that $\sigma_L^{\tilde{L}}$ is the supervenience mapping from L to $\tilde{L}$. The property of having the same image under $\sigma_L^{\tilde{L}}$ forms an equivalence relation on L and the corresponding set of equivalence classes forms a partition of L, that is a set of non-empty, pairwise disjoint, and jointly exhaustive subsets of L. For a value $\tilde{v} \in \tilde{L}$, I will denote the set of its supervenience bases (determinates) in L by $(\sigma_L^{\tilde{L}})^{-1}(\tilde{v}) := \{v \in L : \sigma_L^{\tilde{L}}(v) = \tilde{v}\}$. Since the value $\tilde{v}$ usually indicates the level of consideration, I will usually simplify $(\sigma_L^{\tilde{L}})^{-1}(\tilde{v})$ to $\sigma_L^{-1}(\tilde{v})$. Similarly, I will denote the set of fundamental determinates of $\tilde{v}$ (the determinates of $\tilde{v}$ in the fundamental value set L_0) by $\sigma_0^{-1}(\tilde{v})$.

The idea behind the framework of a causal SVSL-model is of course, that we are not restricted anymore to ascribe only one value set to a variable V in a causal model $\mathcal{M}$. We can instead ascribe several value sets to V, each of which lies on a different supervenience level. Since the distinct value sets that are ascribed to V are arranged in an SVSL, the exact supervenience relationships between the values on the different levels are explicitly represented by the supervenience mappings in the SVSL. A classical causal model with only one value set for each variable is a special case of a causal SVSL-model, since the SVSL of each variable can also entail only one level with the identity supervenience mapping. Since exogenous variables are devoid of any representational meaning, it does not make much sense to ascribe exogenous variables multiple levels of value sets. It is therefore reasonable to demand that in any causal SVSL-model with the signature $(\mathcal{U}, \mathcal{V}, \mathcal{R}_{Lev})$, $\mathcal{R}_{Lev}$ assigns only single level SVSLs (SVSLs that contain only one level) to all $U \in \mathcal{U}$. Exogenous variables will accordingly have only one (fundamental) level. Just as in classical causal models, the context $\vec{u}$ for a causal SVSL-model will therefore denote the (fundamental) values of the exogenous variables.

The framework of causal SVSL-models violates, at least in a certain way, the condition of Mutual Exclusivity. A variable V can take on several values at the same time in a causal SVSL-setting $(\mathcal{M}, \vec{u})$. But it cannot take on several values on one and the same level $L \in \mathcal{L}_V$, since each level, that is each single value set of V, still satisfies Mutual Exclusivity. Instead V takes on exactly one value on each of its levels. Which values of the different levels are compatible with each other is governed by the supervenience functions in the following way:

Axiom 1: If variable V has the value v on level $L \in \mathcal{L}_V$ in $(\mathcal{M}, \vec{u})$, then for every level $\tilde{L} \in \mathcal{L}_V$ with $\tilde{L} > L$: V has the value $\sigma_L^{\tilde{L}}(v)$ on $\tilde{L}$ in $(\mathcal{M}, \vec{u})$.

Axiom 2: If a variable V has the value $\tilde{v}$ on level $\tilde{L} \in \mathcal{L}_V$ in $(\mathcal{M}, \vec{u})$, then for every level L with $L < \tilde{L}$: $\sum_{v \in \sigma_L^{-1}(\tilde{v})} \mathcal{P}(V = v) = 1$.

According to Axiom 1, if V takes on value v on some level $L \in \mathcal{L}_V$ in $(\mathcal{M}, \vec{u})$, then v uniquely determines the values that V takes on at all higher levels. This corresponds to the feature of *Upwards Enheritance* that we have assumed to hold for any supervenience relation under consideration. According to Axiom 2, if V takes on a value $\tilde{v}$ at some level $\tilde{L} \in \mathcal{L}_V$, then $\tilde{v}$ does not uniquely determine the values that V takes on at lower levels. But it ensures that on every lower level, V cannot take on a value that is not a determinate of $\tilde{v}$. This corresponds to the feature of *Requisite and Unique Determination* that we have assumed to hold for any supervenience relation under consideration.

Structural equations in causal SVSL-models work just like in classical causal models. A structural equation F_V tells us, which values the variable V takes on in dependence of the values of certain other variables in the model, which are the parents of V. In the following, I will only consider examples in which the fundamental values of the parents of a variable V uniquely determine the fundamental value of V, which means that a complete context, that is the values of all exogenous variables, uniquely determines a complete causal setting, that is the values of all variables on all levels.[12]

Interventions in causal SVSL-models also work analogously to interventions in classical causal models. If we intervene to set a variable X in a causal model $\mathcal{M}$ to value x on level $L \in \mathcal{L}_X$, then we disrupt the causal influence of all the parents of X on X. Formally, this means that we replace the structural equation F_X in $\mathcal{M}$ by the assignment $X = x$, while all other structural equations in the model remain intact. The resulting causal model is denoted by $\mathcal{M}_{do(X=x)}$. The only thing that we additionally have to consider is the fulfillment of the axioms 1 and 2, which means that we have to adjust the values of X on all the other levels in $\mathcal{L}_X$ such that they are

[12] In general, it may be the case, though, that the fundamental values of V's parents do not uniquely determine the fundamental value of V, but only some higher level value of V. In such a causal SVSL-model, we might have knowdledge about the exact values of all exogenous variables and, consequently, about the exact values of all endogenous variables on some higher levels, while we do not have complete knowledge about the values of the endogenous variables on certain lower levels.

compatible with the value x on L. It is in principle possible to allow 'higher-level' interventions of the form $do(X = \tilde{x})$ with $\tilde{x}$ being a value of X on some non-fundamental level. Such an intervention uniquely determines the values of X on all levels $\hat{L}$ with $\hat{L} > \tilde{L}$, while it does not uniquely determine the values of X on all levels L with $L < \tilde{L}$. But in the following I will, for the sake of simplicity, only consider interventions that uniquely determine the fundamental values of the variables intervened upon.

After having extended the causal model framework to causal SVSL-models, we can now also extend our formal language. First, we can now differentiate between two types of primitive events: fundamental primitive events of the form $X = x$ in which x is a value from the fundamental level $L_0 \in \mathcal{L}_X$, and non-fundamental primitive events of the form $X = \tilde{x}$, in which $\tilde{x}$ is a value from some non-fundamental level $\tilde{L} \in \mathcal{L}_X$. I will use the tilde-symbol over the value ($\tilde{x}$) to denote that the value is from a non-fundamental supervenience level. I will use bold type ($\mathbf{x}$) to indicate that the value may be from any value set level of the variable, either fundamental or non-fundamental. If there is no tilde-symbol over the value and it is not in bold type, then the value is supposed to be from the fundamental value set level of the variable. As in classical causal models, I will say $(\mathcal{M}, \vec{u}) \models X = \mathbf{x}$ if and only if X has the value $\mathbf{x}$ in $(\mathcal{M}, \vec{u})$.[13] Further, $(\mathcal{M}, \vec{u}) \models [X \leftarrow \mathbf{x}]\phi$ if and only if $(\mathcal{M}_{do(X=x)}, \vec{u}) \models \phi$ for all $x \in {\sigma_0}^{-1}(\mathbf{x})$. With $\vec{X} = \vec{\mathbf{x}}$ being shorthand for a conjunction $X_1 = \mathbf{x}_1 \wedge ... \wedge X_n = \mathbf{x}_n$ we will say: $(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{\mathbf{x}}]\phi$ if and only if $(\mathcal{M}_{do(\vec{X}=\vec{x})}, \vec{u}) \models \phi$ for all $\vec{x} \in {\sigma_0}^{-1}(\vec{\mathbf{x}})$, where $\vec{X} = \vec{x}$ with $\vec{x} \in {\sigma_0}^{-1}(\vec{\mathbf{x}})$ is a conjunction $X_1 = x_1 \wedge ... \wedge X_n = x_n$ with $x_1 \in {\sigma_0}^{-1}(\mathbf{x}_1)$, ..., $x_n \in {\sigma_0}^{-1}(\mathbf{x}_n)$.

Before reconsidering the definitions of actual and strong actual causation in the framework of causal SVSL-models, let us briefly consider an example to make the abstract formalism of a causal SVSL-model a bit more vivid. So, how can we extend the classical causal model $\mathcal{M}^S$, that describes Sophie's colour guided pecking behaviour, into a causal SVSL-model? First, in $\mathcal{M}^S$ the variable P can take on two different values: 1, representing that Sophie pecks, and 0, representing that Sophie does'nt peck. Since we are not interested in a more specific or more abstract description of Sophie's pecking behaviour, I will use a single-level SVSL for P with $\mathcal{L}_P = \{L_0 = \{1, 0\}\}$ and with $\mathcal{S}_P$ containing only the identity supervenience mapping from L_0 into L_0. Variable C is more interesting, though. In the classical version of $\mathcal{M}^S$, C's value set only contains the mutually exclusive colour shades *crimson, scarlet, maroon, rose, indigo, azure, ...*. In the SVSL-version of $\mathcal{M}^S$ (SVSL-$\mathcal{M}^S$), we can now ascribe several value sets on different supervenience levels to C, that are organized in an SVSL. We can, for example, use the following SVSL:

- $\mathcal{L}_C = \{L_0, L_1, L_2\}$, with:
  - $L_2 = \{$*warm colour, cold colour*$\}$
  - $L_1 = \{$*red, blue, green, yellow*$\}$
  - $L_0 = \{$*crimson, scarlet, maroon, rose, indigo, azure, ...*$\}$
- $\mathcal{S}_C$ is the closure under S1 and S2 of $\{\sigma_0^1 : L_0 \to L_1, \sigma_1^2 : L_1 \to L_2\}$, with

[13] If one prefers a semantics that only refers to the fundamental values in a causal model, we can alternatively say: $(\mathcal{M}, \vec{u}) \models X = \mathbf{x}$ if and only if $\sum_{x_i \in {\sigma_0}^{-1}(\mathbf{x})} \mathcal{P}(X = x_i) = 1$.

- $(\sigma_0^1)^{-1}(\textit{red}) = \{\textit{crimson, scarlet, maroon, rose}\}$,
  $(\sigma_0^1)^{-1}(\textit{blue}) = \{\textit{indigo, azure, navy, cyan}\}$,
  $\ldots.$
- $(\sigma_1^2)^{-1}(\textit{warm colour}) = \{\textit{red, yellow}\}$, $(\sigma_1^2)^{-1}(\textit{cold colour}) = \{\textit{blue, green}\}$.

To describe the causal relationship between C and P in SVSL-$\mathcal{M}^S$, we can simply use the following structural equation:

- $P = 1$ iff $C = \textit{red}$.

This structural equation is sufficient for capturing the counterfactual dependencies between C and P on all supervenience levels. Due to the explicit representation of the supervenience relations between the values of C in SVSL-$\mathcal{M}^S$, we know, for example, that $[C \leftarrow \textit{scarlet}]P = 1$ is true in any setting for SVSL-$\mathcal{M}^S$, since any intervention to set C to *scarlet* will also set C to *red*, which then yields $P = 1$ according to the structural equation for P. $[C \leftarrow \textit{indigo}]P = 0$ is, for example, also true in any setting for SVSL-$\mathcal{M}^S$, since any intervention to set C to *indigo* will also set C to *blue*, which then yields $P = 0$ according to the structural equation for P. $[C \leftarrow \textit{warm colour}]P = 1$, on the other hand, is false in any setting for SVSL-$\mathcal{M}^S$, since there are interventions that set C to *warm colour* by setting C to *yellow*, which will yield $P = 0$ according to the structural equation for P. But $[C \leftarrow \textit{warm colour}]P = 0$ is also false in any setting for SVSL-$\mathcal{M}^S$, since there are interventions that set C to *warm colour* by setting C to *red*, which will yield $P = 1$ according to the structural equation for P.[14] The causal diagram for SVSL-$\mathcal{M}^S$ is still the same as for the classical version $\mathcal{M}^S$, since it employs just the same variables.

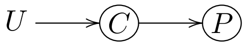


Figure 9.7: Causal diagram for SVSL-$\mathcal{M}^S$.

Let us now consider an additional example with several supervenience levels for both the cause and the effect variable. Imagine that Pete is very poor in keeping track of how much he drinks when he is out with his friends. He therefore made it a habit to measure his own blood alcohol concentration before heading home to decide, if he can safely ride his bike or if he should rather take a cab. For every bottle of beer that Pete drinks, and Pete only drinks beer, his blood alcohol concentration (BAC) increases by a value of 0.03%. We can represent the amount of beer that Pete drinks on a given night by the variable B, for which we will use the following SVSL:

- $\mathcal{L}_B = \{L_0, L_1, L_2\}$, with:
  - $L_2 = \{\textit{less than two beers}, \textit{at least two beers}\}$

[14] The fact that we can use a single structural equation that adequately captures the counterfactual dependencies between C and P on all supervenience levels is a crucial advantage of the SVSL-approach over the two approaches presented in section 9.4. It adequately captures the intuition that the counterfactual dependencies on the different supervenience levels are all associated and are ultimately grounded in only one causal mechanism.

- $L_1 = \{$*drinks beer, does not drink beer*$\}$
- $L_0 = \{0, 1, 2, ..., 20\}$[15]

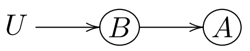
- $\mathcal{S}_B$ is the closure under S1 and S2 of $\{\sigma_0^1 : L_0 \to L_1, \sigma_0^2 : L_0 \to L_2\}$, with
  - $(\sigma_0^1)^{-1}(\textit{drinks beer}) = \{1, 2, ..., 20\}$,
    $(\sigma_0^1)^{-1}(\textit{does not drink beer}) = \{0\}$
  - $(\sigma_0^2)^{-1}(\textit{less than two beers}) = \{0, 1\}$
    $(\sigma_0^2)^{-1}(\textit{at least two beers}) = \{2, ..., 20\}$.

Another variable, A, represents Pete's BAC, for which we will use the following SVSL:

- $\mathcal{L}_A = \{L_0, L_1, L_2\}$, with:
  - $L_2 = \{$*legally allowed to drive, legally not allowed to drive*$\}$
  - $L_1 = \{$*completely sober, alcohol in blood*$\}$
  - $L_0 = \{0, 0.01, 0.02, 0.03, ..., 0.6\}$[16]

- $\mathcal{S}_A$ is the closure under S1 and S2 of $\{\sigma_0^1 : L_0 \to L_1, \sigma_0^2 : L_0 \to L_2\}$, with
  - $(\sigma_0^1)^{-1}(\textit{completely sober}) = \{0\}$,
    $(\sigma_0^1)^{-1}(\textit{alcohol in blood}) = \{0.01, 0.02, 0.03, ..., 0.6\}$
  - $(\sigma_0^2)^{-1}(\textit{legally allowed to drive}) = \{0, 0.01, 0.02, 0.03, 0.04, 0.05\}$[17]
    $(\sigma_0^2)^{-1}(\textit{legally not allowed to drive}) = \{0.06, 0.07, ..., 0.6\}$.

We can use the structural equation and the causal diagram shown in figure 9.8.[18]



- $F_A : A_0 := B_0 \times 0.03$

Figure 9.8: Causal diagram and structural equation for SVSL-$\mathcal{M}^P$.

Here again, the single structural equation F_A is sufficient for capturing the counterfactual dependencies between A and B on all supervenience levels. Due to the explicit representation of the supervenience relations in SVSL-$\mathcal{M}^P$, we know, for example, that $[B \leftarrow \textit{does not drink beer}]A = \textit{completely sober}$ is true in any setting for SVSL-$\mathcal{M}^P$, because any intervention that sets B to *does not drink beer* will also set B to 0 at L_0, which will yield the value 0 for A at L_0 according to the structural equation for A. But $A = 0$ at L_0 is a determinate of $A = \textit{completely sober}$.

[15] The values in L_0 represent the numbers of bottles of beer that Pete might drink on the given night. We do not go any further than the value *20*, because we consider it to be physically impossible for Pete to drink more than 20 bottles of beer on a given night.

[16] The values in L_0 represent Pete's BAC in percent. We do not go any further than the value 0.6, because we consider it to be physically impossible for Pete to have a BAC higher than 0.6%.

[17] We imagine that the legal limit to drive is a BAC of 0.06 in Pete's home state.

[18] The indices on the variables in the structural equation F_A indicate that it is the value of A at A's level L_0 that is determined by the value of B at B's level L_0.

$[B \leftarrow \textit{at least two beers}]A = \textit{legally not allowed to drive}$, for example, is also true in any setting for SVSL-$\mathcal{M}^P$, because any intervention that sets B to *at least two beers* will set B to a value b at L_0 with $b \geq 2$. According to the structural equation for A, this will yield a value a for A at L_0 with $a \geq 0.06$, which will therefore be a determinate of *legally not allowed to drive.*

Let us now adapt our definitions of actual and strong actual causation to the framework of causal SVSL-models. Since the core ingredients of these definitions are counterfactuals of the form $[\vec{X} \leftarrow \vec{\mathbf{x}}]Y = \mathbf{y}$ and since we have already explicated how such counterfactuals are evaluated in causal SVSL-models, both definitions can be adopted without many modifications:

Actual Causation in Causal SVSL-Models. *$\vec{X} = \vec{\mathbf{x}}$ is an 'actual cause' of $Y = \mathbf{y}$ in the causal SVSL-setting $(\mathcal{M}, \vec{u})$ if and only if the following conditions hold:*

AC1 $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{\mathbf{x}} \wedge Y = \mathbf{y}$.

AC2 *There is a partition $(\vec{Z}, \vec{W})$ of the endogenous variables $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$, some alternative value $\vec{\mathbf{x}}'$ of $\vec{X}$ with $\sigma_0^{-1}(\vec{\mathbf{x}}) \cap \sigma_0^{-1}(\vec{\mathbf{x}'}) = \varnothing$, some alternative value $\mathbf{y}'$ of Y with $\sigma_0^{-1}(\mathbf{y}) \cap \sigma_0^{-1}(\mathbf{y}') = \varnothing$ and some setting $\vec{w}$ such that if $(\mathcal{M}, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then:*

(a) $(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{\mathbf{x}'}, \vec{W} \leftarrow \vec{w}]Y = \mathbf{y}'$.

(b) $(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{\mathbf{x}}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]Y = \mathbf{y}$ *for all $\vec{W}' \subseteq \vec{W}$ and for all $\vec{Z}' \subseteq \vec{Z}$.*

AC3 *$\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{\mathbf{x}}'$ (with $\vec{\mathbf{x}}'$ being the restriction of $\vec{\mathbf{x}}$ to the variables in $\vec{X}'$) satisfies conditions (AC1) and (AC2).*

The only two crucial differences to the definition of actual causation in classical causal models is that we allow the values $\vec{\mathbf{x}}$ and $\mathbf{y}$ to be from any value set level of the respective variables and that we forbid that $\vec{\mathbf{x}}$ and $\mathbf{y}$ have any overlaps with their respective alternatives $\vec{\mathbf{x}'}$ and $\mathbf{y}'$ on the respective fundamental supervenience levels in the given causal SVSL-model.[19] With just the same adaptations, we can adopt the definition of strong actual causation into the framework of causal SVSL-models:

Strong Actual Causation in Causal SVSL-Models. *$\vec{X} = \vec{\mathbf{x}}$ is a 'strong actual cause' of $Y = \mathbf{y}$ in the causal SVSL-setting $(\mathcal{M}, \vec{u})$ if and only if the following conditions hold:*

SAC1 $(\mathcal{M}, \vec{u}) \models \vec{X} = \vec{\mathbf{x}} \wedge Y = \mathbf{y}$.

SAC2 There is a partition $(\vec{Z}, \vec{W})$ of the endogenous variables $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$, some alternative value $\vec{\mathbf{x}}'$ of $\vec{X}$ with $\sigma_0^{-1}(\vec{\mathbf{x}}) \cap \sigma_0^{-1}(\vec{\mathbf{x}'}) = \varnothing$, some alternative value $\mathbf{y}'$ of Y with $\sigma_0^{-1}(\mathbf{y}) \cap \sigma_0^{-1}(\mathbf{y}') = \varnothing$ and some setting $\vec{w}$ such that if $(\mathcal{M}, \vec{u}) \models Z = z^*$ for all $Z \in \vec{Z}$, then:

(a) $(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{\mathbf{x}'}, \vec{W} \leftarrow \vec{w}]Y = \mathbf{y}'$.

(b) $(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{\mathbf{x}}, \vec{W}' \leftarrow \vec{w}, \vec{Z}' \leftarrow \vec{z}^*]Y = \mathbf{y}$ *for all $\vec{W}' \subseteq \vec{W}$ and for all $\vec{Z}' \subseteq \vec{Z}$.*

SAC3 Let $\vec{V}$ be all variables in $\mathcal{V}$ that are not in any causal path from a variable in $\vec{X}$ to Y that is active in $(\mathcal{M}, \vec{u})$, then: $(\mathcal{M}, \vec{u}) \models [\vec{X} \leftarrow \vec{\mathbf{x}}, \vec{V} \leftarrow \vec{v}]\phi$ for all settings $\vec{v}$ of $\vec{V}$.

[19] This last demand is not necessary in classical SVSL-models, since it is automatically satisfied due to Mutual Exclusivity.

SAC4 $\vec{X}$ *is minimal; there is no strict subset* $\vec{X}'$ *of* $\vec{X}$ *such that* $\vec{X}' = \vec{x}'$ *(with* $\vec{x}'$ *being the restriction of* $\vec{x}$ *to the variables in* $\vec{X}'$*) satisfies conditions SAC1, SAC2 and SAC3.*

With these definitions at our disposal, we can now evaluate causal claims on several supervenience levels in a causal SVSL-model. Imagine, for example, that we have a setting for SVSL-$\mathcal{M}^S$, in which the object is scarlet. A simple application of the definitions just presented will then enable us to acknowledge the truth of the two following causal claims in the given causal setting: 'the object being scarlet is a (strong) actual cause of Sophie's pecking' and 'the object being red is a (strong) actual cause of Sophie's pecking'. But even though the object has a warm colour in the given situation, the causal claim 'the object having a warm colour is a (strong) actual cause of Sophie's pecking' turns out to be false, since condition AC2(b) (SAC2(b)) is violated: We can set the colour of the object on a warm colour, namely yellow, without thereby yielding Sophie to peck. Applying our definitions of causation to SVSL-$\mathcal{M}^P$ in a causal setting, in which Pete drinks 3 beers, will, for example, enable us to acknowledge the truth of the following causal claims in the given causal setting: 'Pete's consumption of 3 beers is a (strong) actual cause of him having a BAC of 0.09', 'Pete's consumption of 3 beers is a (strong) actual cause of him having alcohol in his blood', or: 'Pete's consumption of more than 2 beers is a (strong) actual cause of him not being legally allowed to drive'.[20] This illustrates, that causal claims on several supervenience levels can now be evaluated relative to a single causal SVSL-model, which means that the expressivity problem is solved.

9.6 Summary

In the present chapter, I have pointed to a crucial shortcoming of interventionist accounts of causation that are based on classical causal models: Intuitively adequate causal claims on different supervenience levels about a single causal scenario cannot be evaluated in a single causal model. I have dubbed this problem the *expressivity problem.* After having argued that the expressivity problem is indeed a vice and not a virtue, since causal exclusion arguments do not apply to interventionist concepts of causation, I have shortly discussed two existing approaches to address the expressivity problem. I have argued that both approaches have the significant downside that supervenience relationships are not explicitly modeled and that, consequently, intuitively connected causal claims on different supervenience levels are treated as being completely unrelated. I have therefore put forward a new approach to address the expressivity problem, which consists in the extension of classical causal models to causal SVSL-models. The framework of a causal SVSL-model is able to explicilty represent supervenience relationships between different events and it enables the evaluation of causal claims on several supervenience levels in a single causal model.

[20] I omit the proofs for the causal claims mentioned here, since in all these cases we have a simple counterfactual dependence of the effect on the cause, such that we can use $\vec{W} = \varnothing$ in condition AC2 (SAC2).

Chapter 10

Proportionality

10.1 Introduction

In the previous chapter, I have argued that causal exclusion arguments do not apply to 'thin' concepts of causation, like the interventionist concepts of actual and strong actual causation. 'Thin' causal relationships between two variables can exist on several supervenience levels without yielding a counterintutive overdetermination of the effect. The object being scarlet can be a strong actual cause of Sophie's pecking, while the object being red is also a strong actual cause of Sophie's pecking, even though Sophie's pecking is clearly not causally overdetermined by these two events. As a consequence, we also end up with two different causal explanations of Sophie's pecking: We can explain Sophie's pecking either by citing the scarlet colour of the object shown to her or by citing the red colour of the object shown to her. The account of causal explanation that I have developed in this dissertation is built on the concept of strong actual causation. So, as soon as I acknowledge that an effect can have several strong actual causes on different supervenience levels, I automatically acknowledge that the effect can have several causal explanations on different supervenience levels. A natural question therefore arises: Given an explanandum ϕ, is there some level of supervenience for a causal explanation of ϕ that is superior to other levels?

Several authors have argued that there is. According to these authors, a causal explanation of an explanandum ϕ should be *proportional* to ϕ, in the sense that it should be neither too abstract nor too specific for ϕ. In the present chapter, I aim to explore the idea of proportionality by employing the framework of causal SVSL-models.

10.2 Proportional Actual Causes in Causal SVSL-Models

One of the key proponents of proportionality is Yablo (1992). According to Yablo, a cause or a causal explanation of an explanandum ϕ can be either too abstract or too specific.[1] It is too abstract, if some of its determinates are no causes of ϕ after all. Take the Sophie-scenario as an example. 'The object has a warm colour' is too abstract for a successful causal explanation

[1] Like Yablo (1992), I will treat proportionality first and foremost as a feature of causes. A causal explanation is proportional to its explanandum ϕ if all the causes that are cited in the causal explanation are proportional to ϕ.

of Sophie's pecking. Although 'The object has a warm colour' has a determinate that is able to cause Sophie's pecking, namely 'the object is red', it also has a determinate that is unable to cause Sophie's pecking, namely 'the object is yellow'. The more specific event 'the object is red' therefore fares essentially better as a causal explanation of Sophie's pecking. But if we try to explain Sophie's pecking by saying that the object is scarlet, then our explanation has become overly specific. Citing the red colour of the object would have been enough to explain Sophie's pecking. According to Yablo (1992), being overly specific means to convey irrelevant information, which is confusing at best and wrong at worst.[2] This is why 'the object is red' also fares better as an explanation of Sophie's pecking than 'the object is scarlet'.

The idea of proportionality can be summarized like this: A proportional cause is specific enough, but only as specific as required. Based on this idea, Yablo has segmented the constraint of proportionality into two sub-conditions: The first condition, called *Enough*, makes sure that the cause is specific enough, while the second condition, called *Required*, ensures that the cause is only as specific as required. A cause is proportional relative a given effect, if and only if it is enough and required for the effect. A first semiformal explication of the two sub-conditions of proportionality for actual causes can be formulated like this:[3]

(Enough) An actual cause $X = \mathbf{x}$ of $Y = \mathbf{y}$ is *enough* for $Y = \mathbf{y}$ if and only if every determinate $X = \mathbf{x}^+$ of $X = \mathbf{x}$ is also an actual cause of $Y = \mathbf{y}$.

(Required) An actual cause $X = \mathbf{x}$ of $Y = \mathbf{y}$ is *required* for $Y = \mathbf{y}$ if and only if there is no more abstract determinable $X = \mathbf{x}^-$ of $X = \mathbf{x}$ that is an actual cause of $Y = \mathbf{y}$, that is enough for $Y = \mathbf{y}$.

Now, how can these conditions be made precise in our formal framework of causal SVSL-models? First and foremost, it should be pointed out that, according to our definition of actual causation in causal SVSL-models, every actual cause of an effect $Y = \mathbf{y}$ is enough for $Y = \mathbf{y}$. This becomes obvious, as soon as we consider the semantics for higher level formulas in causal SVSL-models. For example, $(\mathcal{M}, \vec{u}) \models X = \mathbf{x}$ if and only if $(\mathcal{M}, \vec{u}) \models X = x_i$ for every $x_i \in \sigma_0^{-1}(\mathbf{x})$. Also, $(\mathcal{M}, \vec{u}) \models [X \leftarrow \mathbf{x}]Y = \mathbf{y}$ if and only if $(\mathcal{M}, \vec{u}) \models [X \leftarrow x_i]Y = \mathbf{y}$ for every $x_i \in \sigma_0^{-1}(\mathbf{x})$. From this it quickly follows that if $X = \mathbf{x}$ is an actual cause of $Y = \mathbf{y}$, so is $X = x_i$ for every $x_i \in \sigma_0^{-1}(\mathbf{x})$. Any actual cause $X = \mathbf{x}$ of an effect $Y = \mathbf{y}$ in a causal SVSL-model is therefore proportional relative to $Y = \mathbf{y}$ as soon as it is required for $Y = \mathbf{y}$. I therefore propose the following definition of proportionality of actual causes in causal SVSL-models:

Proportional Actual Cause (Prop-AC). *Let $X = \mathbf{x}$ be an actual cause of $Y = \mathbf{y}$ in a causal SVSL-model $(\mathcal{M}, \vec{u})$. $X = \mathbf{x}$ is proportional relative to $Y = \mathbf{y}$ in $(\mathcal{M}, \vec{u})$ if and only if it satisfies the following condition:*

(Req) *There is no value x' of X at X's fundamental value set level L_0 with $x' \notin \sigma_0^{-1}(\mathbf{x})$ such that $X = x'$ would also be an actual cause of $Y = \mathbf{y}$ in $(\mathcal{M}_{do(X=x')}, \vec{u})$.*

[2] See the discussion about causally irrelevant information in chapter 1.

[3] My explication of both conditions deviates in certain respects from Yablo's (1992, p. 276-7) original formulation, since he builds his conditions on a simplified concept of causation. My explication nonetheless retains the essential idea underlying Yablo's proportionality constraint.

It may not be immediately obvious that *Req* reproduces the content of *Required.* To see that this is indeed the case, notice that the fulfillment of *Req* makes sure that it is impossible to find a determinable $X = \mathbf{x}^-$ of $X = \mathbf{x}$ that has both $X = \mathbf{x}$ and some additional fundamental event $X = x'$ as determinates such that $X = \mathbf{x}^-$ is also an actual cause of (and therefore enough for) $Y = \mathbf{y}$.[4]

Notice that our definition of an SVSL allows for a determinable to have only a single determinate at a lower level. If $X = \mathbf{x}$ is the only determinate of the determinable $X = \mathbf{x}^-$ and both are actual causes of the effect, then both will be proportional to the effect, as long as they both satisfy *Req.* It is therefore possible that a given effect has several proportional causes that are related in terms of supervenience.[5]

10.3 Event-Proportionality vs. Variable-Proportionality

There is yet another popular definition of proportionality that I want to mention only briefly. The definition has been put forward by Woodward (2010) and Franklin-Hall (2016), as well as Blanchard (2020), discuss it in some detail. Although Woodward's definition of proportionality is related to Yablo's concept of proportionality, which the definition Prop-AC is supposed to reproduce, it differs in a crucial respect. According to Prop-AC, proportionality is a relation between two events, $X = \mathbf{x}$ and $Y = \mathbf{y}$, in a setting for a causal SVSL-model. But according to Woodward's (2010) definition, proportionality is a relation between two variables, X and Y, in a causal model. The basic idea of Woodward's definition is this: Two variables, X and Y, that stand in a causal relationship to each other, are proportional to each other, if and only if every value of X is able to causally yield one and only one value of Y, if realized by an intervention, and there are no two different values of X that would causally yield the same value of Y, if realized by an intervention.[6] Take $\mathcal{M}^S$ as an example. Variable C is not variable-proportional to variable P, if C's value set is $\{$*crimson, scarlet, maroon, rose, indigo, azure, ...*$\}$, since C has several values, namely *crimson, scarlet, maroon,* and *rose*, that would all causally yield $P = 1$ if realized by an intervention. And it has several values, namely *indigo, azure,* and so on, that would all causally yield $P = 0$ if realized by an intervention. If C's value set is $\{$*red,* $\neg$*red*$\}$, on the other hand, then C is variable-proportional relative to P.[7]

Woodward's concept of variable-proportionality can be helpful for ensuring event-proportinality,

[4] One might wonder, why I do not use the following, rather straightforward, explication of *Required*: There is no determinable $X = \mathbf{x}^-$ of $X = \mathbf{x}$, that has both $X = \mathbf{x}$ and some additional fundamental event $X = x'$ as determinates, such that $X = \mathbf{x}^-$ is also an actual cause of $Y = \mathbf{y}$. The reason is that, according to this explication, the actual cause $X = \mathbf{x}$ of $Y = \mathbf{y}$ is proportional relative to $Y = \mathbf{y}$ as soon as the causal SVSL-model under consideration does not contain any determinable for $\mathbf{x}$. *Req* avoids this kind of model-relativity. But, as we will see in section 10.4, *Req* and the resulting definition of proportionality is still model-relative in several respects. But, as I will argue, there are several well-entrenched rules of causal modeling, which ensure that the remaining model-relativity of our definition of proportionality does not lead to counterintuitive outcomes.

[5] We could, of course, amend our definition of SVSLs and demand that any determinable on a given level L must have multiple realizers on all lower levels. Proportionality would then be restricted to a single supervenience level. But I do not see any reason to impose this additional constraint.

[6] See (Woodward, 2010, p. 298).

[7] In causal SVSL-models, it does, of course, not make much sense to talk about variable-proportionality, since every variable may have several different value set levels. But Woodward's definition of variable-proportionality can easily be adapted to a definition of level-proportionality, according to which proportionality is a relation between two value set levels of two different variables X and Y.

since it holds that, if $X = \mathbf{x}$ is an actual cause of $Y = \mathbf{y}$ and X is variable-proportional relative to Y, then $X = \mathbf{x}$ is event-proportional relative to $Y = \mathbf{y}$. But $X = \mathbf{x}$ can very well be a proportional actual cause of $Y = \mathbf{y}$, even if X and Y are not variable-proportional. Since I am interested in the proportionality of causal explanations and since causal explanations are events, my interest is in the concept of event-proportionality. In the following, I will therefore not discuss the concept of variable-proportionality any further.

10.4 The Model-Relativity of Proportionality

10.4.1 Franklin-Hall's Objection

Franklin-Hall (2016) has put forward an argument according to which the explication of proportionality in the framework of causal models is a hopeless endeavor. Although her criticism is directed at Woodward's (2010) definition of variable-proportionality, it can be applied just as well to our definition Prop-AC. Franklin-Hall's argument can be nicely illustrated with the Sophie-scenario. Take the model SVSL-$\mathcal{M}^S$ and imagine that $C =$ *scarlet*. Prop-AC then accurately recognizes that $C =$ *scarlet* is no proportional actual cause of $P = 1$ in SVSL-$\mathcal{M}^S$, since there are alternative fundamental values of C that, if realized by an intervention, would also be actual causes of $P = 1$ in the given causal setting, namely, for example, *maroon*. Instead, only $C =$ *red* is acknowledged as a proportional actual cause of $P = 1$ in SVSL-$\mathcal{M}^S$. Franklin-Hall now asks us to make a slight change to the causal (SVSL-)model, namely to reduce C's fundamental value set from $L_0 = \{$*crimson, scarlet, maroon, rose, indigo, azure, ...*$\}$ to $L_0' = \{$*scarlet, indigo*$\}$. With L_0' as C's fundamental value set, Prop-AC will identify $C =$ *scarlet* as a proportional actual cause of $P = 1$, if $C =$ *scarlet* holds in the given causal setting. The reason is, that there is now no alternative value of C that, if realized by an intervention, would also be an actual cause of $P = 1$ in the given causal setting. With this example, Franklin-Hall aims to reveal what she considers to be a fatal model-relativity of the formal definition of proportionality: A very simple and seemingly harmless change to the causal (SVSL-)model under consideration and our definition of proportionality yields an entirely different and contradicting result.

Franklin-Hall's argument rightly points out that our definition of proportionality is model-relative. But this does not make the definition futile. It only highlights that, when working with causal (SVSL-)models, we have to make sure that the models we use are adequate representations of the scenarios that we aim to discuss. In the next section, I will argue that, if we are going to abide by certain well-entrenched and well-justified rules of causal modeling, we will see that our definition of proportionality will work just fine and that it accords with common intuitions.

10.4.2 Some Rules of Causal Modeling

The question of what makes a causal model adequate and what rules to follow in the process of causal modeling to make a model adequate is definitely underrepresented in the literature on causal models. But this does not mean that there is a lack of causal modeling rules. On the contrary, as already pointed out in the last chapter, most rules of causal modeling, while well-entrenched and carefully attended to, are just tacitly presupposed in the literature. Only a few authors have taken up the task to make these rules explicit, especially Halpern and Hitchcock

(2010) and Woodward (2016). In this section, I will not come close to providing a comprehensive list of causal modeling rules. Instead, I will only put forward three such rules, which turn out to be helpful for the problem at hand. All three rules have already been described in the literature as fundamental rules of causal modeling and each rule is motivated quite independently from considerations about proportionality. I will argue that, if we abide by these rules and apply our definition of proportionality, as well as all our model-relative definitions of causation, only to causal models that conform to these rules, then the definition will yield reasonable and intuitively adequate results. Now, here is the first rule that I want to consider:[8]

Exhaustivity of Value Sets (EVS). *Any value set of any variable V in a causal model $\mathcal{M}$ should exhaust all seriously possible outcomes.*

The motivation for EVS is this: If we restrict a value set of a variable in a causal model $\mathcal{M}$ by omitting seriously possible outcomes that may exert causal influences on other variables in $\mathcal{M}$, then our causal model not only denies real possibilities about how the represented situation may actually develop, it also omits causal information about the situation at hand. This is exactly what happens in Franklin-Hall's modification of $\mathcal{M}^S$. If we use a causal model with the (fundamental) value set $\{$*scarlet, indigo*$\}$ for variable C, then this model omits the information that there are still other shades of red, which can cause Sophie to peck, while the model also neglects the clearly possible outcome that the object shown to Sophie might take on some colour different from scarlet or indigo. If we would take this model at its word, then the represented situation is such that the object shown to Sophie can only be scarlet or indigo and the only relevant causal information about the situation is that a scarlet object causes Sophie to peck, while an indigo object does not. This is a crucially different causal scenario than the one described by $\mathcal{M}^S$ and we should not be surprised that Prop-AC yields a different result when applied to this modified scenario. The problem with Franklin-Hall's modification of $\mathcal{M}^S$ is not that our definition Prop-AC is inadequate. The problem is rather that Franklin-Hall's causal model is inadequate as a representation of the original Sophie-scenario.

For the second rule that I want to consider, I first have to introduce the concept of an *ambiguous intervention.*[9] Consider a variable X in a causal model $\mathcal{M}$ with x being some value of X. It is crucial for any causal model approach to causation that it is, at least in principle, possible to intervene on X and set it to the value x. Formally, this is done by the do-operator $do(X = x)$. While an intervention $do(X = x)$ has to be quite precise to conform to the demands placed on the do-operator, the operation $do(X = x)$ does rarely ever uniquely determine one distinct real-life action. Typically, an intervention $do(X = x)$ can be implemented through various different actions. Consider again the Sophie-scenario with the value set $\{$*red, blue, green, yellow*$\}$ for C. The intervention $do(C = red)$ can be implemented by, for example, showing Sophie a scarlet object, showing Sophie a crimson object, or by showing Sophie a maroon object. All these different actions are legitimate implementations of the intervention $do(C = red)$, which makes

[8] Woodward (2016) also mentions this guideline: "We also want our variables to take a range of values corresponding to the full range of genuine or serious possibilities that can be exhibited by the system of interest" (Woodward, 2016, p. 1064). A very similar causal modeling rule has also been put forward by Blanchard (2020, cf. p. 649).

[9] The concept is introduced by Spirtes and Scheines (2004), although they use the term 'ambiguous manipulation'.

the intervention, to a certain degree, ambiguous. Now, here is our second rule:[10]

Same Effects for Realizations of Ambiguous Interventions (SERAI). *The effects of an ambiguous (fundamental) intervention* $do(X = x)$ *in a causal (SVSL-) model* $\mathcal{M}$ *should, at least to our best knowledge, not depend on how* $do(X = x)$ *is implemented.*

Notice that SERAI does not prohibit ambiguous interventions per se. Only if it is known that different ways to implement an ambiguous intervention can yield different effects in some setting for the causal model $\mathcal{M}$, $\mathcal{M}$ is considered to be defective. If we would, for example, take $\{$*warm colour, cold colour*$\}$ as the fundamental value set level for C in SVSL-$\mathcal{M}^S$, then we would have a violation of SERAI, since we know that $do(C =$ *warm colour*$)$ can be implemented by two different actions that would yield different effects in SVSL-$\mathcal{M}^S$: Showing Sophie a red object would yield $P = 1$, while showing Sophie a yellow object would yield $P = 0$. As a result, the event $C =$ *warm colour* does not satisfy Yablo's condition *Enough* relative to $P = 1$, since it is not concrete enough such that every determinate of it would be an actual cause of $P = 1$, if realized by an intervention. This is why SVSL-$\mathcal{M}^S$ needs to include more specific values of C to enable the identification of what we intuitively understand to be (strong) actual causes.

Before coming to the third rule of causal modeling, let us first consider another problematic example that has been put forward by Franklin-Hall (2016): Imagine that Sophie not only pecks whenever she sees a red object, but that she also pecks whenever she is provided with food. Additionally, we can electrically stimulate Sophie's cerebellum, which also causes her to peck. Intuitively, this gives us three independent factors that can causally influence Sophie's pecking behaviour. Now, Franklin-Hall invites us to represent this causal scenario by using a causal model with the two variables C and P, where P, as always, has the value set $\{0, 1\}$, and C has the value set $\{$*showing a red object* $\vee$ *providing food* $\vee$ *stimulating the cerebellum*, *not showing a red target* $\wedge$ *not providing food* $\wedge$ *not stimulating the cerebellum*$\}$. The causal relationship between P and C is then described by the following structural equation:

- $P = 1$ iff $C =$ *showing a red object* $\vee$ *providing food* $\vee$ *stimulating the cerebellum*

The causal model proposed by Franklin-Hall does not violate EVS, because the two values of C indeed exhaust all seriously possible outcomes. The model also conforms to SERAI: Even though there are several different ways for implementing the intervention $do(C =$ *showing a red object* $\vee$ *providing food* $\vee$ *stimulating the cerebellum*$)$, all these different interventions will yield the same effect in the given causal model, namely $P = 1$.[11] But applying our definition Prop-AC will yield a highly counterintuitive result in the proposed causal model, since it identifies $C =$ *showing a red object* $\vee$ *providing food* $\vee$ *stimulating the cerebellum* as a proportional actual cause of Sophie's pecking, if C takes on this value in the given causal setting. Intuitively, $C =$ *showing a red object* $\vee$ *providing food* $\vee$ *stimulating the cerebellum* is extremely unproportional relative to $P = 1$, because it is way to abstract to be an adequate causal explanation of Sophie's pecking. So what went wrong? Our next rule provides an answer:

[10] Spirtes and Scheines (2004, cf. p. 833) already consider this rule to be a standard assumption in causal modeling.

[11] The same is true for C's alternative value.

No Lumping (NL). *Outcome spaces that can be manipulated independently of each other (which means that fixing the outcome for one outcome space does not restrict either metaphysically, logically, mathematically, or conceptually the outcome for the other outcome spaces) should not be lumped together in a single variable.*[12]

The causal model proposed by Franklin-Hall is a clear violation of NL. Whether or not Sophie is shown a red object, whether or not Sophie is provided with food, and whether or not Sophie's cerebellum is electrically stimulated, are three distinct outcome spaces that can be manipulated independently of each other. An adequate causal model would therefore use a separate variable for each of these outcome spaces. Notice that, yet again, there is a motivation for NL as a rule of causal modeling that is quite independent from considerations about proportionality. As both Blanchard (2020) and Woodward (2021) point out, a causal model of a given scenario that violates NL is severely restricted in answering, what Woodward calls, 'what-if-things-had-been-different-questions' in comparison to causal models of the same scenario that conform to NL. Just take Franklin-Hall's NL-violating causal model of her Sophie-scenario and imagine that in the actual situation a red object is shown to Sophie, while her cerebellum is not stimulated and no food is provided to her. Franklin-Hall's lumped variable C then takes on the value $C =$ *showing a red object* $\vee$ *providing food* $\vee$ *stimulating the cerebellum*. But the causal model does not give us any information that goes beyond this disjunction. It does not tell us, which or how many of the three disjuncts are actually true in the given situation. This is why the causal model can also not tell us what would happen, if we would only manipulate the colour of the object and make it blue, without doing anything else. An NL-violating model is incapable of informing us about the causal effects of what we intuitively understand to be simple interventions. Using a causal model, in which each of the three outcome spaces gets its own variable, solves this issue.

So, what Franklin-Hall's examples have shown is the following: If we apply our definition of Prop-AC to inadequate causal models, Prop-AC may very well yield counterintuitive results and identify causes as proportional that are either way too abstract or way too specific. But this should not come as a surprise. It is to be expected that model-relative defintions will yield counterintuitive results about a given causal scenario, if we use a causal model that does not adequately describe the scenario. This is why Franklin-Hall's examples do not amount to arguments against the adequacy of Prop-AC. But they may serve as a useful reminder to only use adequate causal models, which includes the conformity to EVS, SERAI, and NL.[13]

10.5 Proportionality: A Necessary Condition for Causation?

By defending the definition Prop-AC against Franklin-Hall's argument, I have also defended the conviction that it is indeed possible to formulate a meaningful condition of proportionality within the framework of causal models. But now that a causal-model-based condition of proportionality has been saved, another question arises: What significance should we ascribe to the condition of proportionality when it comes to interventionist concepts of causation? Does being proportional

[12] Both Blanchard (2020) and Woodward (2021) argue for a rule like NL.

[13] Blanchard (2020) puts forward a very similar response to Franklin-Hall's argument, but his focus lies on Woodward's conception of variable-proportionality.

really make a cause somehow better or more valuable than a non-proportional alternative? Or should it even be acknowledged as a necessary condition for causation per se? In section 10.6, I will explore whether being proportional makes a cause somehow superior to its non-proportional alternatives. In the present section, I want to focus on the question, whether proportionality should be acknowledged as a necessary condition for causation.

10.5.1 Proportionality as a Saviour of Higher-Level Causation

List and Menzies (2009) take proportionality as a necessary condition for a difference-making concept of causation. They do so with a certain objective in mind, which is to save certain higher-level causal claims from Kim's causal exclusion argument. As illustrated in the previous chapter, the causal exclusion argument concludes that causation is restricted to one supervenience level only. Many proponents of causal exclusion assume that it must be the fundamental supervenience level that is causally efficacious and thereby exclude the causal efficacy of events that lie on higher supervenience levels. By taking proportionality as a necessary condition for causation, List and Menzies (2009) agree with the position of causal exclusion: Causation is restricted to one supervenience level only. But they break with the popular assumption that this level must be the fundamental level. According to their account, it is the proportional level instead.

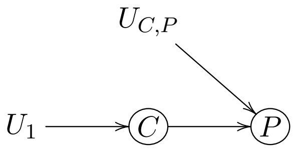
Although I agree with the motivation to safe higher level causal claims, which are often highly intuitive, from Kim's causal exclusion argument, I do not think that we need proportionality as a necessary condition for causation to achieve this. At least not when it comes to interventionist concepts of causation. As argued in the previous chapter, Kim's causal exclusion argument does not apply to interventionist concepts of causation in the first place. The fact that taking proportionality as a necessary condition for causation does itself amount to a form of causal exclusion, rather speaks against this move. As argued in the previous chapter, non-proportional causal claims are often highly intuitive. This is an intuition that is not only held by some philosophers, like Shapiro and Sober (2012) or Woodward (2021), but that seems to be widely shared in general, as empirical studies by Blanchard et al. (forthcoming) suggest.

Nonetheless, there are still some further arguments that can be made for taking proportionality as a necessary condition for causation. For the discussion of these arguments, it is helpful to divide the condition of proportionality into its sub-conditions: *Enough* and *Required*. As pointed out in section 10.1, the interventionist account of causation, that I have adopted in this dissertation, accepts *Enough* as a necessary condition for causation. I will start with discussing whether this position is reasonable. Subsequently, I will deal with the question of whether *Required* should be a necessary condition for causation.

10.5.2 Why Should *Enough* Be Necessary for Causation?

In many situations we speak of causation, even though the cause does not guarantee its effect with complete certainty. The account of actual causation defended in this dissertation allows for an event C to be an actual cause of an effect E, even if C's causal power on E is lower than 1. But according to the same account of actual causation, any actual cause C must be *enough* for its effect E. This may seem overly harsh, since not being *enough* for the effect is just one way of not being sufficient for the effect. So, if an account of actual causation already admits certain causes that are not sufficient for its effects, why does'nt it also admit causes that are not *enough* for its effects?

For addressing this question, it is important to note that the condition of *Enough*, which our definition of actual causation conforms to, is relativized to the causal knowledge represented in the given causal model. It may well be the case that an event $X = x$ is acknowledged as an actual cause of an event $Y = y$ (and therefore as being *enough* for $Y = y$) in a causal setting $(\mathcal{M}, \vec{u})$, while $X = x$ is not *enough* for $Y = y$ from an objective or omniscient perspective. Imagine that showing Sophie a red object does not cause Sophie to peck with a maximal causal power of 1. Instead, a randomized controlled trial yielded the result that showing Sophie a red object only causes her to peck with a causal power of 0.7. We do not know why the redness of the object is not generally sufficient for Sophie's pecking. As far as we know, this could be a case of genuine indeterminism. We can use the causal SVSL-model $\mathcal{M}^S$ to represent this scenario, if we incorporate the following amendments: We have to introduce an error-term $U_{C,P}$, which represents whether $C = red$ successfully causally produces $P = 1$, given that $C = red$ is the case. We can then use the SEM illustrated in figure 10.1 to represent the scenario.



- $C := U_1$
- $P = 1$ iff $C = red$ $\wedge U_{C,P} = 1$.

Figure 10.1: Causal SVSL-model $\mathcal{M}^{US}$ - the unreliable Sophie scenario.

In a setting with $U_{C,P} = U_1 = 1$, our definition of actual causation acknowledges $C = red$ as an actual cause of $P = 1$, even though we know that $C = red$ is, in general, not sufficient for $P = 1$. But now imagine that we come to learn that in the randomized controlled trial, which determined the causal power of $C = red$ on $P = 1$, the following happened: In 70% of the cases, in which Sophie was shown a red object, the experimenter showed her a scarlet object, which reliably yielded Sophie to peck, while in the other 30%, the experimenter showed Sophie an object that has some other shade of red, which reliably yielded Sophie not to peck. The structural equation '$P = 1$ iff $C = red$ $\wedge U_{C,P} = 1$' does not adequately represent this newly gained causal knowledge, while the following structural equation does:

- $P = 1$ iff $C = scarlet$.

With this new structural equation, our account of actual causation will not admit $C = red$ as an actual cause of $P = 1$ anymore, since we now know that $C = red$ is not enough for $P = 1$. What has changed is not the objective causal relationship between C and P, but only our knowledge about the causal relationship. It is this knowledge-relative condition of *Enough* that our causal model approach takes to be necessary for causation and that can be paraphrased like this: A cause has to be enough for its effect according to our best causal knowledge. This shows that our account of actual causation often acknowledges causes that are, objectively, not enough for their effects. But it does so only as long as we do not have the knowledge that enables the formulation of more specific causes that are enough for their effects.

But why should we adopt the knowledge-relative version of *Enough* as a necessary condition for causation? Coming back to our example, in which Sophie only pecks in reaction to scarlet objects: why should we not be allowed to also admit $C = \textit{red}$, in addition to $C = \textit{scarlet}$, as a cause of $P = 1$? As we have seen in our example, using a more abstract description of a cause, a description that is not enough for the effect of interest, means to add further uncertainty about whether the cited cause is able to yield the effect. Whether we use the causal relationship for a causal explanation, for a prediction, or for making a decision, the additional uncertainty always makes the cause that is not enough for the effect inferior to its *Enough*-satisfying alternative. This is why rejecting the constantly inferior candidate entirely as a genuine cause is a reasonable act of parsimony.

10.5.3 Should *Required* Be Necessary for Causation?

Should we also adopt *Required*, and therefore Proportionality as a whole, as a necessary condition for causation? Here is the popular and often silently accepted claim that could be used as an argument for this position: A cause that violates *Required* includes information that is causally irrelevant for its effect.[14] If this claim is true, there is a strong similarity between unproportionally specific causal claims and causal claims that include causally irrelevant conjuncts. In chapter 1, I have already argued that causally irrelevant conjuncts are deadly for causal claims and causal explanations. Claiming that $X = x \wedge Z = z$ causes or causally explains $Y = y$ entails the claim that both $X = x$ and $Z = z$ are causally relevant to $Y = y$. This motivated the minimality conditions AC3 and SAC4 as necessary constraints for actual and strong actual causation, respectively. But if unproportionally specific causal claims include causally irrelevant information for their effects, just like causal claims with causally irrelevant conjuncts do, shouldn't we then adopt *Required* as a necessary condition for causation as well? I think the answer should be yes. But the crucial question is, whether unproportionally specific causes really do contain causally irrelevant information for their effects. I think that this popular claim is actually quite questionable when considered thoroughly.

Clearly, whether unproportionally specific causes contain causally irrelevant information for their effects depends crucially on how we define causal relevance, and it is not so obvious how to do so. Despite supporting the claim that unproportionally specific causes convey causally irrelevant information in (Woodward, 2010), Woodward (2021) provides a definition of causal relevance, according to which unproportionally specific causes end up being entirely causally relevant for their effects. Ignoring cases of preemption and overdetermination, the definition can

[14] Woodward (2010), Yablo (1992), Blanchard (2020), and Strevens (2008) all make this claim. For Strevens the causal irrelevance claim forms a crucial motivation for his kairetic account of causal explanation that aims to eliminate causally irrelevant information from causal explanations by abstracting away from unproportionally specific descriptions of the cited causes. Yablo (1992) and Woodward (2010) do not draw the conclusion from the causal irrelevance claim, that proportionality should be necessary for causation. Yablo (1992) states: “Without claiming that proportionality is strictly necessary for causation, it seems clear that faced with a choice between two candidate causes, normally the more proportional candidate is to be preferred“ (Yablo, 1992, p. 277). And Woodward (2021) writes that the “satisfaction of a proportionality requirement should not be regarded as a necessary condition for a causal claim to be true” (Woodward, 2021, p. 244).

be formulated like this: an event $X = x$ is causally relevant for another event $Y = y$ in a causal setting $(\mathcal{M}, \vec{u})$ with $(\mathcal{M}, \vec{u}) \models X = x \wedge Y = y$ if and only if there is an alternative value x' of X such that $(\mathcal{M}_{do(X=x')}, \vec{u}) \models Y = y'$ with $y' \neq y$.[15] The definition gives adequate results for classical examples of causal irrelevance. For example, hexing the table salt is correctly identified as being causally irrelevant to whether the salt dissolves in water or not, and taking birth-control pills is correctly identified as being causally irrelevant to whether a man gets pregnant or not. But unproportionally specific causes satisfy Woodward's relevance condition. Taking $\mathcal{M}^S$ as an example, the object being scarlet is identified as being causally relevant to Sophie's pecking, because there is an intervention that changes the colour of the object from scarlet to some alternative colour, which leads to a change in Sophie's pecking behaviour.

One could object that Woodward's definition of causal relevance is overly permissive. His condition can be strenghtened by requiring that $X = x$ is only causally relevant for $Y = y$ if every possible intervention that changes the value of X also leads to a change of the value of Y. This stronger condition of causal relevance would indeed ensure that unproportionally specific causes are identified as causally irrelevant for their effects. But is there any reason to prefer the stronger condition of causal relevance over the weaker one proposed by Woodward (2021)? Or, alternatively, is there any reason to prefer the weaker condition of causal relevance over the stronger one? The results of the studies done by Blanchard et al. (forthcoming) seem to speak for the weaker definition of causal relevance: People often accept unproportionally specific causal claims, which, at least, indicates that they also ascribe some causal relevance to unproportionally specific causes. I share the intution that denying the object being scarlet all causal relevance for Sophie's pecking behaviour seems to be overly harsh. But at the same time, it is just as intuitive to say that it would have been enough to know that the object was red to predict or to explain Sophie's pecking behaviour. Knowing the specific shade of red is indeed irrelevant. Or, to put it differently, learning that the object is scarlet is only relevant insofar that it implies that the object is red and anything above this implied information is irrelevant. This intuition can be accommodated by another concept introduced by Woodward (2021), namely the concept of *conditional irrelevance*, which he explicates like this: $X = x$ is conditionally irrelevant to $Y = y$, given $Z = z$, if and only if: if one intervenes to fix Z at value z, then further variations in X due to interventions consistent with Z being fixed at z will not change the value of Y.[16] If we now opt for the weaker definition of causal relevance, we can say that, while unproportionally specific causes are causally relevant for their effects, they are conditionally irrelevant to their effects, given their proportional determinables. I think that this account best fits the conflicting intuitions about the causal relevance of unproportionally specific causes. But the crucial question is then this: Should $X = x$ not be acknowledged as a cause of $Y = y$, only because there is some $Z = z$ such that $X = x$ is conditionally irrelevant to $Y = y$, given $Z = z$?

Note that for any indirect cause $X = x$ of the effect $Y = y$ with intermediary $Z = z$, $X = x$

[15] See (Woodward, 2021, p. 242). The formulation of causal relevance given here slightly differs from Woodward's formulation, since Woodward (2021) applies his definition of causal relevance to variables and not to events. The formulation given here is the straightforward implementation of the same idea for events.

[16] See Woodward (2021, p. 254). Here again, Woodward's original formulation applies to variables and not to events.

is conditionally irrelevant to $Y = y$, given $Z = z$. But we would clearly not say that indirect causes are no genuine causes of their effects, just because they are conditionally irrelevant to their effects. Conditional irrelevance per se is therefore clearly not sufficient for denying an event the status of being a cause. But imagine we would say '$X = x \wedge Z = z$ is a cause of $Y = y$', while $X = x$ is conditionally irrelevant to $Y = y$, given $Z = z$. In that case, the causal claim appears to be defective. By not just citing $X = x$ as a cause of $Y = y$, but also $Z = z$, which screens off $X = x$ from $Y = y$ in one and the same causal claim, we have now indeed made a claim that entails causally irrelevant information. I do not see any reason why this should be different for unproportionally specific causes: It does not make sense to ban causal claims that cite unproportionally specific causes, only because they are conditionally irrelevant to their effects, given more abstract determinables. But it makes sense to ban a causal claim that includes both, an unproportionally specific cause $X = x$ of $Y = y$ and a more abstract cause $Z = z$ of $Y = y$ with $Z = z$ being a determinable of $X = x$, since $X = x$ is conditionally irrelevant to $Y = y$, given $Z = z$. According to this position, 'the object is red and the object is scarlet' is not acknowledged as a cause of Sophie's pecking. But 'the object is scarlet' is acknowledged as a genuine cause of Sophie's pecking, even though it is not *required* and therefore conditionally irrelevant for Sophie's pecking.[17]

In summary, I consider it well justified that *Enough* is taken as a necessary condition for causation. But the same cannot be said for *Required*. The only reasonable argument that can be made for taking *Required* as a necessary condition for causation is based on the claim that unproportionally specific claims contain causally irrelevant information for the effect in question. But I have shown that this claim does not stand up to closer scrutiny. Still, even if there is no good reason to take *Required*, and therefore proportionality, as a necessary condition for causation, it could still be a feature that makes causes especially valuable or even superior to their unproportional alternatives. I will now explore whether this is indeed the case. As Woodward (2014) points out, it depends crucially on ones goals, which characteristics of causes amount to valuable features. Since the focus of this dissertation is on causal explanations, I am going to focus on the question of whether proportionality is a valuable feature of a cause that is employed for a causal explanation.

10.6 Proportionality as a Virtue for Causal Explanations

10.6.1 The Value of Being as Abstract as Possible

Unification of Causal Information

As Blanchard (2020) points out, causes that are required for and therefore proportional to their effects contain more causal information than their more specific alternatives. Proportional causes provide, to use a term coined by Hitchcock and Woodward (2003), more "resources for answering what-if-things-had-been-different questions" [w-questions] by making explicit what the value of the explanandum variable depends upon" (Hitchcock and Woodward, 2003, p. 190). For example, if the object shown to Sophie is scarlet and Sophie pecks, then saying that the object being

[17] Notice that our definitions of actual and strong actual causation in causal SVSL-models concur with this assessment.

scarlet caused Sophie to peck, answers less w-questions than saying that the object being red caused Sophie to peck. The last statement makes explicit that Sophie would also have pecked, if the object would have had any other shade of red. Citing the proportional cause therefore unveils more factors that are able to causally yield the effect. This makes a proportional description of a cause a more convenient and unifying way of storing and conveying causal knowledge than a more specific description of the cause. As pointed out in section 8.2.1, unification is a popular and frequently suggested explanatory virtue,[18] even in the context of causal explanations. Strevens (2008), for example, considers the feature of unification to be a crucial achievement of his kairetic account of causal explanation, which explicitly aims at providing proportional causes as explanations for a given explanandum. However, unification accounts of explanation have long faced the challenge of providing an appropriate formal explication of unification. A formal explication of proportionality can clearly be of help in addressing this challenge, at least when it comes to causal explanations.

Increasing the Likeliness of the Explanans

When it comes to inferences to the best explanation (IBE), proportional causes provide another advantage over unproportionally specific causes. A proportional cause is simply more likely to have happened than a more specific determinate of it. Given that Sophie has pecked and we do not know why, then the hypothesis that Sophie was shown a red object is more likely to be true than the hypothesis that Sophie was shown a scarlet object, since the last hypothesis entails the first. Inferring a proportional cause is therefore a more reliable ineference than inferring a more specific alternative in the process of IBE.

10.6.2 The Value of Being Specific

Mechanism-Sensitivity

While the abstractness of proportional causes yields certain advantages, there is also a counteracting virtue of being highly specific. An important purpose of a causal explanation is to provide a certain amount of control over the explanandum. Learning what causes a given effect can provide us with the ability to change it by manipulating the mechanism by which the cause yields the effect. As the following example shows, being unproportionally specific can be advantageous for this purpose.

Imagine that Pete is at a wedding party where there is a photobox that enables the guests to shoot selfies. When shooting a selfie, a guest has three options: By pressing the first button, the photobox takes a photo and sends it to the phones of all guests at the location via WiFi exactly one minute after the photo was taken. By pressing the second button, the photobox takes a photo and sends it to the phones of all guests at the location via Bluetooth exactly one minute after the photo was taken. And by pressing the third button, the photobox takes a photo and does not send it to anyone. The process of sending a selfie to a phone is not flawless in both cases. In the first case, the transmission of a photo to the phone of a guest may fail, when either the photobox or the phone is not connected to the Internet. In the second case, the transmission

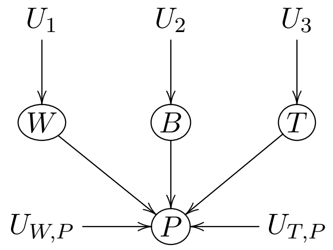
[18] See, for example, (Friedman, 1974), (Kitcher, 1976), (Schurz and Lambert, 1994).

may fail, when either the photobox or the phone does not have a bluetooth connection. Now the following happens. Together with his friend Tom, Pete has made a selfie at the photobox. Tom pressed the first button. But Pete looks terrible on the photo. Since he has a crush on Lena, another guest at the party, he really does not want the photobox to send the photo to Lena's phone. Pete suddenly has a very high interest in which button Tom has pressed, since he wants to know, how he might stop the transmission of the embarrassing photo to Lena's phone. And he only has one minute to succeed. He cannot tinker with the photobox itself, because its settings cannot be accessed by any guest. He also has no access to Lena's phone. So, he sees only two options, he either shuts down the WiFi network of the location by finding the router and pulling its plug and thereby disconnects the photobox from the internet, or he activates his Bluetooth jammer, which he coincidentally has in his car in front of the wedding location, to jam the photobox' bluetooth connection. But it is impossible for him to do both in just one minute.

We can represent the causal scenario by a causal SVSL-model that contains the following variables: Variable B represents which button of the photobox is pressed. Its fundamental value set contains values that represent the three different buttons: $L_0 = \{1, 2, 3\}$. But we can also introduce a higher value set level $L_1 = \{\mathit{send}, \neg \mathit{send}\}$, in which *send* represents that the photobox takes a photo and somehow sends it to Lena's phone and $\neg$*send* represents that the photobox just takes a photo without sending it. We therefore have the following supervenience relations: $\sigma_0^{-1}(\mathit{send}) = \{1, 2\}$ and $\sigma_0^{-1}(\neg \mathit{send}) = \{3\}$. Variable P represents whether Lena receives the photo on her phone. Its only value set is $L_0 = \{0, 1\}$, with 0 representing that she does not receive the photo and 1 representing that she does receive the photo. The variable W represents whether the photobox is connected to the internet. Its only value set is $L_0 = \{0, 1\}$, with 1 representing that it is and 0 representing that it is not connected to the internet. Variable T represents whether the photobox has a bluetooth connection. Here again, its only value set is $L_0 = \{0, 1\}$, with 1 representing that it has and 0 representing that it is does not have a bluetooth connection. The error-term $U_{W,P}$ with the value set $\{0, 1\}$ is supposed to capture any further disturbances in the internet connection between the photobox and Lena's phone, including whether Lena's phone has a working internet connection. And the error-term $U_{T,P}$ with the value set $\{0, 1\}$ is supposed to capture any further disturbances in the bluetooth connection between the photobox and Lena's phone, including whether Lena's phone has a working bluetooth connection. The causal SVSL-model $\mathcal{M}^P$ that represents the described scenario is illustrated in figure 10.2.

Let us assume that the following probability distribution $\mathcal{P}$ represents Pete's epistemic state: $\mathcal{P}(U_1 = 1) = \mathcal{P}(U_3 = 1) = 1$, since we assume that Pete knows that the photobox has an intact internet and bluetooth connection. We assume that Pete is ignorant about which button Tom has pressed. Pete therefore assigns the same probability to all three possibilities: $\mathcal{P}(U_2 = 1) = \mathcal{P}(U_2 = 2) = \mathcal{P}(U_2 = 3) = 1/3$. And we assume that Pete ascribes a high probability, say 0.8, to the fact that Lena's phone would receive the photo, if the photobox sends it via WiFi and a high probability, say 0.7, to the fact that Lena's phone would receive the photo, if the photobox sends it via bluetooth. This gives us: $\mathcal{P}(U_{W,P} = 1) = 0.8$ and $\mathcal{P}(U_{T,P} = 1) = 0.7$.

With $\vec{u} = (u_1, u_2, u_3, u_{W,P}, u_{T,P})$, $B = 1 \wedge W = 1$ is a strong actual cause of $P = 1$ in



- $W := U_1$
- $B := U_2$
- $T := U_3$
- $P = 1$ iff $(B = 1 \wedge W = 1 \wedge U_{W,P} = 1) \vee (B = 2 \wedge T = 1 \wedge U_{T,P} = 1)$

Figure 10.2: Causal SVSL-model $\mathcal{M}^P$ - the photobox scenario.

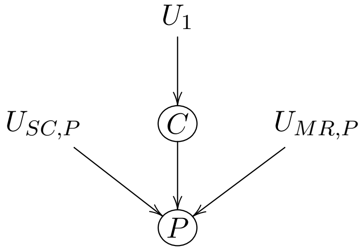
the following contexts with a non-zero probability: $(1,1,1,1,0)$ and $(1,1,1,1,1)$. This makes $B = 1 \wedge W = 1$ a potential explanation of $P = 1$. $B = 2 \wedge T = 1$ is a strong actual cause of $P = 1$ in the following contexts with a non-zero probability: $(1,2,1,0,1)$ and $(1,2,1,1,1)$. This makes $B = 2 \wedge T = 1$ a potential explanation of $P = 1$. But $B = \textit{send} \wedge W = 1 \wedge T = 1$ is also a strong actual cause of $P = 1$ in the following contexts with a non-zero probability: $(1,1,1,1,1)$ and $(1,2,1,1,1)$. This makes $B = \textit{send} \wedge W = 1 \wedge T = 1$ a potential explanation of $P = 1$.[19]

The potential explanation $B = \textit{send} \wedge W = 1 \wedge T = 1$ is clearly not specific enough for Pete's purpose. He wants to know how to prevent $P = 1$, whether he should get his bluetooth jammer and thereby disable the causal process from $B = 2 \wedge T = 1$ to $P = 1$ or whether he should shut down the WiFi and thereby disable the causal process from $B = 1 \wedge W = 1$ to $P = 1$. Learning that $B = \textit{send} \wedge W = 1 \wedge T = 1$ does not help him with this task in any way. For this, he needs to know the fundamental value of B, that is whether B has the value 1 or the value 2. Learning only that $B = \textit{send}$ is too abstract for Pete's purpose, since it does not enable him to identify which of the two distinct mechanisms, by which $B = \textit{send}$ is able to causally yield $P = 1$, was initiated. But, according to our definition of proportionality of actual causes, $B = \textit{send}$ is a proportional actual cause of $P = 1$ in both contexts $(1,1,1,1,1)$ and $(1,2,1,1,1)$, while $B = 1$ is an unproportionally specific actual cause of $P = 1$ in $(1,1,1,1,1)$ and $B = 2$ is an unproportionally specific actual cause of $P = 1$ in $(1,2,1,1,1)$. In both cases, there is an alternative value of B that, if realized by an intervention, would also be an actual cause of $P = 1$. We therefore have a clear example, where using an unproportionally specific description of the cause is more valuable for the purpose of control over a given explanandum than using a proportional description of the cause.

Consider another example: We have the causal SVSL-model $\mathcal{M}^S$ of the Sophie scenario with the system of value set level for variable C as introduced in section 9.5. Only now, we face different causal relationships between the specific values of C and P. Yet again, only red colour shades are able to cause Sophie to peck, but different shades of red do so with different

[19] Note that we have the following extensive causal explanations: $P = 1$, *because* $B = 1 \wedge W = 1$. $P = 1$, *because* $B = 2 \wedge T = 1$. And $P = 1$, *because* $B = \textit{send} \wedge W = 1 \wedge T = 1$.

causal powers. Results of randomized controlled trials have shown that scarlet and crimson seem to have exactly the same effect on whether Sophie pecks or not and both yielded Sophie's pecking with a causal power of 0.8. Maroon and rose also truned out to have exactly the same effect on whether Sophie pecks or not, but both yielded Sophie's pecking with a causal power of 0.5. The causal SVSL-model that represents this adapted Sophie-scenario therefore needs to be complemented with two error terms: $U_{SC,P}$, which represents whether the object being scarlet or crimson would successfully produce Sophie's pecking, if the object would indeed be scarlet or crimson, and $U_{MR,P}$, which represents whether the object being maroon or rose would successfully produce Sophie's pecking, if the object would indeed be maroon or rose. The resulting causal SVSL-model $\mathcal{M}^{S'}$ is depicted in figure 10.3.



- $C := U_1$
- $P = 1$ iff $(C = \textit{scarlet} \wedge U_{SC,P} = 1) \vee (C = \textit{crimson} \wedge U_{SC,P} = 1) \vee (C = \textit{maroon} \wedge U_{MR,P} = 1) \vee (C = \textit{rose} \wedge U_{MR,P} = 1)$

Figure 10.3: Causal SVSL-model $\mathcal{M}^{S'}$ - the adapted Sophie scenario.

Now imagine a situation, in which we come to learn that Sophie has pecked, but we do not yet know, which value C has taken on. $C = \textit{red}$ is a strong actual cause of $P = 1$ in all contexts with $U_{SC,P} = 1$ and $U_{MR,P} = 1$, which makes $C = \textit{red}$ a potential explanation of $P = 1$. Since $C = \textit{red}$ is additionally a maximal potential strong actual cause of $P = 1$ relative to $\mathcal{M}^{S'}$, we get the extensive causal explanation '$P = 1$ *because* $C = \textit{red}$'. According to Prop-AC, $C = \textit{red}$ is a proportional cause relative to $P = 1$. But according to our account of explanatory power, the potential explanation $C = \textit{red}$ is not specific enough to have a specific explanatory power. This is why $C = \textit{red}$ is intuitively too abstract as an explanation of $P = 1$ in the given situation. We know that different shades of red have different causal powers on Sophie's pecking. So trying to explain Sophie's pecking just by stating that the object is red, provokes the natural question: Yes, sure, but which shade of red? As long as this question is unanswered, we are not in a position to evaluate the power of the explanation that is given to us.

We therefore have an example, where an unproportionally specific description of the cause is necessary for providing causal explanations that can be evaluated and compared in terms of their intrinsic explanatory power.

Stability

Being *mechanism-sensitive*, that is, being specific enough to differentiate between different mechanisms by which a cause is able to yield the explanandum, does not only provide a higher level of control over the explanandum. It is also connected to another virtue of causation that Woodward accentuates: Stability.[20] Consider the causal claim that $B = \textit{send}$ is an actual cause of $P = 1$ in the Photobox-scenario. The truth of the claim depends on several background conditions: $W = 1$, $P = 1$, $U_{WP} = 1$, and $U_{TP} = 1$ all have to be true for '$B = \textit{send}$ is an actual cause of $P = 1$' to be true. The more specific causal claim that $B = 1$ is an actual cause of $P = 1$ depends on less background conditions. Only $W = 1$ and $U_{WP} = 1$ have to be true for '$B = 1$ is an actual cause of $P = 1$' to be true. This is what makes the second causal claim more stable than the first: It is less susceptible to changes in background conditions and, in cases of uncertainty about which background conditions hold, therefore also more likely to be true, given that the cause-candidate itself is assumed to be the case. Since a causal explanation of an explanandum is only correct if the explanation candidate itself is true and the explanation candidate actually causes the explanandum, a higher stability of a causal relationship makes a potential explanation, ceteris paribus, more likely to be correct.

10.7 Proportional Causal Explanations

In the last section, I have identified two counteracting virtues of causal explanations. First, a virtue of being as abstract as possible to yield a simple and unified description of the potential causes of the explanandum and to make the explanans itself more likely. And secondly, a virtue of being specific enough to differentiate between different mechanisms by which the explanandum can be brought about, which provides us with a higher amount of control over the explanandum and with more stable causal relationships. In this section, I am going to propose an explication of proportionality that reconciles both virtues in the following way: A causal explanation is proportional to its explanandum if and only if it is as abstract as possible while still being specific enough to be mechanism-sensitive. Since all our definitions of causal explanations are based on the concept of potential strong actual causation, it suffices to explicate this concept of proportionality for potential strong actual causes.[21] A causal explanation, may it be a potential, actual, partial, explicitly complete, or extensive explanation, is then proportional if the potential strong actual causes used for the explanation are proportional.

Just like Yablo's explication of proportionality, my explication will consist of two conditions: The first condition is there to ensure that the proportional potential strong actual cause is specific enough to be mechanism-sensitive relative to the given explanandum. The second condition then ensures that the proportional potential strong actual cause, that is mechanism-sensitive relative to the given explanandum, is as abstract as possible.

But first some terminology: Remember that an event $X = \mathbf{x}$ is a determinate of an event $X = \mathbf{x}^-$ in a causal SVSL-model $\mathcal{M}$ if and only if there is a supervenience mapping σ in $\mathcal{S}_X$ such that $\sigma(\mathbf{x}) = \mathbf{x}^-$. Notice that σ may also be the identity mapping, which means that every event $X = \mathbf{x}$ is a determinate of itself in every causal SVSL-model. Now, for a conjunction $\vec{X} = \vec{\mathbf{x}}$ we will say: $\vec{X} = \vec{\mathbf{x}}$ is a determinate of $\vec{X} = \vec{\mathbf{x}}^-$ if and only if for every conjunct $X_i = \mathbf{x_i}$ in $\vec{X} = \vec{\mathbf{x}}$: $X_i = \mathbf{x_i}$ is a determinate of $X_i = \mathbf{x_i}^-$, where $X_i = \mathbf{x_i}^-$ is a conjunct in $\vec{X} = \vec{\mathbf{x}}^-$. For example,

[20] See, for example, (Woodward, 2006) and (Woodward, 2010).

[21] Remember that a potential strong actual cause $\vec{X} = \vec{x}$ of an effect ϕ in a probabilistic causal model $(\mathcal{M}, \mathcal{P})$ is an event that has a non-zero probability of being a strong actual cause of ϕ, that is: $\mathcal{P}(\vec{X} = \vec{x} \rightarrowtail^{\mathcal{M}} \phi) > 0$.

$B = 1 \wedge W = 1 \wedge T = 1$ is a determinate of $B =$ *send* $\wedge W = 1 \wedge T = 1$ in $\mathcal{M}^P$, because $B = 1$ is a determinate of $B =$ *send*, $W = 1$ is a determinate of $W = 1$, and $T = 1$ is a determinate of $T = 1$ in $\mathcal{M}^P$. We can now define:

Mechanism-Sensitive (MS). *A potential strong actual cause $\vec{X} = \vec{\mathbf{x}}$ of ϕ in a causal SVSL-model $(\mathcal{M}, \mathcal{P})$ is 'mechanism-sensitive' relative to ϕ if and only if:*

(MS1) For every logically possible context $\vec{u}$ for $\mathcal{M}$, it holds that: If $\vec{X} = \vec{\mathbf{x}}$ is a strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$, then every determinate $\vec{X} = \vec{\mathbf{x}}^+$ of $\vec{X} = \vec{\mathbf{x}}$ is a strong actual cause of ϕ in $(\mathcal{M}_{do(\vec{X}=\vec{\mathbf{x}}^+)}, \vec{u})$.

(MS2) For every logically possible context $\vec{u}$ for $\mathcal{M}$, it holds that: If a determinate $\vec{X} = \vec{\mathbf{x}}^+$ of $\vec{X} = \vec{\mathbf{x}}$ is a strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$, then $\vec{X} = \vec{\mathbf{x}}$ is also a strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$.

As a first example, consider $B =$ *send* $\wedge W = 1 \wedge T = 1$ in $(\mathcal{M}^P, \mathcal{P})$. As pointed out, $B =$ *send* $\wedge W = 1 \wedge T = 1$ is a potential strong actual cause of $P = 1$ in $(\mathcal{M}^P, \mathcal{P})$, because it is a strong actual cause of $P = 1$ in $(\mathcal{M}^P, (1, 1, 1, 1, 1))$ and $(\mathcal{M}^P, (1, 2, 1, 1, 1))$. But it is not mechanism-sensitive relative to $P = 1$, since $B = 1 \wedge W = 1 \wedge T = 1$ is a determinate of $B =$ *send* $\wedge W = 1 \wedge T = 1$, but $B = 1 \wedge W = 1 \wedge T = 1$ is, for example, no strong actual cause of $P = 1$ in $(\mathcal{M}^P_{do(B=1 \wedge W=1 \wedge T=1)}, (1, 1, 1, 1, 1))$.[22] We therefore have a violation of condition MS1.

Next, consider the potential strong actual cause $C =$ *red* of $P = 1$ in $(\mathcal{M}^{S'}, \mathcal{P})$. $C =$ *red* is not mechanism-sensitive relative to $P = 1$, because there is a determinate of $C =$ *red*, for example $C =$ *scarlet*, that is a strong actual cause of $P = 1$ in a causal setting with $C =$ *scarlet*, $U_{SC,P} = 1$, and $U_{MR,P} = 0$. But $C =$ *red* is no strong actual cause of $P = 1$ in the very same setting, since $[C \leftarrow \textit{red}]P = 1$ is false in this setting: If C is set to, for example, *maroon*, it would not yield Sophie to peck. We therefore have a violation of condition MS2.

Now, the second condition for our new concept of proportionality must ensure that a mechanism-sensitive potential strong actual cause is as abstract as possible. For this, we first introduce the concept of a twin of a potential strong actual cause:

Twin of a Potential Strong Actual Cause. *Let $\vec{X} = \vec{\mathbf{x}}$ be a potential strong actual cause of ϕ in the causal SVSL-model $(\mathcal{M}, \mathcal{P})$. $\vec{X} = \vec{\mathbf{x}}'$ is a twin of $\vec{X} = \vec{\mathbf{x}}$ relative to ϕ in $\mathcal{M}$ if and only if:*

(Twin1) There is at least one variable X_i in $\vec{X}$ such that X_i has a value $\mathbf{x}'_i$ in $\vec{X} = \vec{\mathbf{x}}'$ with $\mathbf{x}'_i \cap \mathbf{x}_i = \varnothing$, where $\mathbf{x}_i$ is the value of X_i in $\vec{X} = \vec{x}$.

(Twin2) For every logically possible context $\vec{u}$ for $\mathcal{M}$, it holds that: If $\vec{X} = \vec{\mathbf{x}}$ is a strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$, then $\vec{X} = \vec{\mathbf{x}}'$ is a strong actual cause of ϕ in $(\mathcal{M}_{do(\vec{X}=\vec{\mathbf{x}}')}, \vec{u})$.

(Twin3) For every logically possible context $\vec{u}$ for $\mathcal{M}$, it holds that: If $\vec{X} = \vec{\mathbf{x}}'$ is a strong actual cause of ϕ in $(\mathcal{M}, \vec{u})$, then $\vec{X} = \vec{\mathbf{x}}$ is a strong actual cause of ϕ in $(\mathcal{M}_{do(\vec{X}=\vec{\mathbf{x}})}, \vec{u})$.

[22] This is due to the minimality condition SAC4 of strong actual causation: $B = 1 \wedge W = 1$ is a strong actual cause of $P = 1$ in $(\mathcal{M}^P_{do(B=1 \wedge W=1 \wedge T=1)}, (1, 1, 1, 1, 1))$. So, according to SAC4, $B = 1 \wedge W = 1 \wedge T = 1$ is no strong actual cause of $P = 1$ in $(\mathcal{M}^P_{do(B=1 \wedge W=1 \wedge T=1)}, (1, 1, 1, 1, 1))$.

For example, in $\mathcal{M}^{S'}$, $C = scarlet$ is a potential strong actual cause of $P = 1$ and $C = crimson$ is a twin of $C = scarlet$. Both events are a strong actual cause of $P = 1$ in exactly the same causal settings. We can now define:

Comprising. *A potential strong actual cause $\vec{X} = \vec{\mathbf{x}}$ of ϕ in $(\mathcal{M}, \mathcal{P})$ is 'comprising' relative to ϕ if and only if there is no twin of $\vec{X} = \vec{\mathbf{x}}$ relative to ϕ in $\mathcal{M}$.*

MS-Proportional Potential Strong Actual Cause. *A potential strong actual cause $\vec{X} = \vec{\mathbf{x}}$ of ϕ in $(\mathcal{M}, \mathcal{P})$ is 'MS-proportional' relative to ϕ if and only if $\vec{X} = \vec{\mathbf{x}}$ is mechanism-sensitive and comprising relative to ϕ.*

For example, in the Photobox-scenario $B = send \wedge W = 1 \wedge T = 1$ is no MS-proportional potential strong actual cause of $P = 1$ in $(\mathcal{M}^P, \mathcal{P})$, because, as illustrated in section 10.6.2, $B = send \wedge W = 1 \wedge T = 1$ is not mechanism-sensitive relative to $P = 1$. $B = 1 \wedge W = 1$, on the other hand, is an MS-proportional potential strong actual cause of $P = 1$ in $(\mathcal{M}^P, \mathcal{P})$, because it is mechanism-sensitive and comprising. The same holds for $B = 2 \wedge T = 1$.

In the Sophie-scenario as described by $\mathcal{M}^{S'}$, our definition of MS-proportionality yields the result that $C = red$ is an unproportionally abstract potential strong actual cause of $P = 1$, since it is not mechanism-sensitive. But $C = scarlet$, $C = maroon$, $C = rose$, and $C = crimson$ are all unproportionally specific, since none of them is comprising. $C = warm\ colour$, on the other hand, is too abstract, for the simple reason that $C = warm\ colour$ is no potential strong actual cause of $P = 1$ in the first place. We therefore have no MS-proportional level in the given causal SVSL-model. For this we would have to introduce another level, in which we have one determinable that encompasses only *scarlet* and *crimson* as determinates and another determinable that encompasses only *maroon* and *rose* as determinates.

As already pointed out in the introduction to this chapter, all the concepts of causal explanation, that I have defined in this dissertation, are based on the concept of potential strong actual causation. We can therefore say, that a causal explanation is MS-proportional relative to its explanandum if and only if the potential strong actual cause that constitutes the explanation is MS-proportional relative to the explanandum. Before I conclude, I would like to point out that I do not want to claim that the definition of MS-proportionality, that I have just proposed, always picks out the most valuable or useful supervenience level. It clearly depends on our exact purposes and goals, whether a more abstract or a more specific causal explanation may turn out to be more useful. But I think, that the here defined level of MS-proportionality is a reasonable tradeoff between the different virtues identified in section 10.6, especially when it comes to causal explanations that are supposed to serve as hypotheses in an Inference to the Best Explanation. This is why I propose that, unless there are very specific reasons in a given situation to use a different level of specificity for a causal explanation, we should by default reach for the causal explanations of a given explanandum ϕ that are MS-proportional to ϕ in the sense just defined.

10.8 Summary

I have started this chapter by providing an explication of Yablo's (1992) concept of proportionality that is supposed to apply to the concept of actual causation in the newly introduced framework of causal SVSL-models. I have defended the very idea of explicating proportionality in the causal model framework against Franklin-Hall's argument by pointing out that a model-relative concept of proportionality is unproblematic as long as we follow some well-entrenched rules of causal modeling. After having argued that there is no reason to take the condition *Required* as a necessary condition for causation, I have illustrated that abstraction and specificity are two counteracting virtues when it comes to causal explanations. I have ultimately proposed a new concept of proportionality that seems to be a reasonable tradoff between these different virtues.

Conclusion

I would like to briefly summarize the main results of this dissertation, point to some of the problems that I have left untouched, and make some suggestions on how future research may build upon the results that I have presented in the preceding ten chapters.

The main goal of this dissertation was to provide a foundation for a formalization of IBE. With regard to this goal it has produced three major outcomes.

First, I have shown that the HP-definition of explanation, despite its promising features, often leads to unintuitive and unwarranted results. I have therefore proposed amendments to overcome these inadequacies and thereby put forward several new definitions of different, but related concepts of causal explanation, including the concepts of *potential*, *actual*, *correct*, *partial*, *ambivalent partial*, *explicitly complete*, and *extensive causal explanation.* Just like the HP-definition of explanation, all these concepts are explicated in the framework of causal models and they all treat causal explanations as being relative to a given epistemic state. I have illustrated how the definitions can be applied in contexts with probabilistic causal relationships.

Secondly, I have demonstrated that several popular Bayesian measures of explanatory power are incapable of measuring an explanation's intrinsic ability to reduce the degree of surprise about a given explanandum. I have therefore put forward a new method for determining explanatory power that is able to do so when it comes to causal explanations.

Finally, I have argued that abstraction and specificity are two counteracting virtues for causal explanations and I have illustrated that Yablo's (1992) characterization of proportionality does not reliably pick the most suitable level of abstraction for causal explanations. I have therefore developed a new explication of proportionality, namely MS-proportionality, which is generally better suited for picking the optimal level of abstraction for a potential causal explanation.

Beyond the three main objectives, there are many new valuable insights that we have gained along the way. Chapter 2 throws light on the differences between Causal Bayes Nets and Structural Equation Models when it comes to the representation of probabilistic causal relationships and it illustrates that both formal frameworks yield different interpretations of actual causation. Chapters 4 and 5 provide an expansion of Cheng's (1997) Power PC theory and thereby yield more generally applicable measures of intrinsic generative causal power and intrinsic preventive power. Chapters 4 to 6 enhance our understanding of two causal concepts that, according to several philosophers as well as psychologists,[23] play a crucial role in our causal understanding of the world: Causal production and prevention. Besides providing a new explication of prevention, which strongly builds on Dowe's (2000) counterfactual theory of prevention, Chapter

[23] See, for example,(Dowe, 2000), (Hall, 2004), (Illari, 2011), (Cheng, 1997), and (Walsh and Sloman, 2011).

6 offers a first sketch of an account that analyzes the concept of causal production without the need to assume some sort of spatio-temporal connection between cause and effect. Instead, the account only employs the interventionist concept of strong actual causation and the distinction between default and deviant states in a given causal setting. Finally, Chapter 9 provides a formal extension of the causal model framework that enables to explicitly represent supervenience relationships in a causal model. This framework of causal SVSL-models not only allows to evaluate causal claims on different supervenience levels relative to a single causal model, it also enables to explicate proportionality constraints in the causal model framework as shown in chapter 10.

Concerning the three main results of my dissertation, which are supposed to lay a foundation for a formalization of IBE, there are certain restrictions that should be kept in mind. First and foremost, I have not provided an account of explanation per se, which would need to encompass all kinds of explanation that may exist. What I have put forward in this dissertation is an account of causal explanation of token events which does not make any claim whatsoever about what other kinds of explanation there might be and how they might be formally explicated. This also means that my account does not contradict or preclude any account of non-causal explanation, like unificationist accounts, or accounts of mathematical,[24] metaphysical,[25] or renormalization-group explanations.[26] Nonetheless, it is an interesting question, a question that I have left untouched, how different kinds of explanation are related and whether there is some common ingredient, some kind of explanation essence, so to speak. There are recent accounts that aim to provide such a comprehensive characterization of explanation. For example, de Regt and Dieks (2005) and de Regt (2017) consider understanding to be the essence of explanation. Reutlinger (2016, 2017) and Reutlinger et al. (2020), on the other hand, claim that it is counterfactual dependencies that make both causal and non-causal explanations explanatory. Both theories fit very well with the account of causal explanation developed in this dissertation, since the account assumes that any causal explanation is able to reduce an agent's surprise about the explanandum and to thereby increase the agent's understanding of the factual world and it assumes the explanandum to be counterfactually (de facto) dependent on the explanation.

While causal-model-based accounts of causation and explanation typically have the advantage of being precisely explicated formally, they seem to share a weakness, that can also be held against the account developed in this dissertation. It is common practice in the causal model literature that the definitions explicated in the causal model framework are only tested on rather simple 'toy models' that can only provide highly simplified reflections of real-life scenarios.[27] My dissertation is no exception. It was my endeavour to consider causal scenarios that are well known to be problematic for many accounts of causation, like scenarios that include overdetermination, back-up causes, preemption, omissions, prevention, and double-prevention. I also came up with new scenarios that proved to be problematic for the HP-definition of explanation and to which I applied my new definitions. It is nevertheless true that all these examples are highly simplified models of real-life scenarios.[28] A critic might therefore point out that even if

[24] See, for example, (Steiner, 1978) and (Lange, 2013).

[25] See, for example, (Schaffer, 2016), (Schaffer, 2017).

[26] See, for example, (Batterman, 2000). Furthermore, Reutlinger (2017) provides a concise overview over recent literature on non-causal explanations.

[27] See (Glymour et al., 2010) for a related criticism of causal-model-based definitions of actual causation.

[28] As pointed out in section 1.3.2, one of these simplifications is the assumption that for any considered epistemic

all these causal-model-based definitions function properly when applied to simple toy examples, it remains entirely unclear how well they will work if confronted with the complexity of the real world. I am sympathetic to this concern. But I do not think that it renders the work done so far futile. The first obvious reply is clearly this: We have to start somewhere. And it is sensible to first test our definitions on rather simple scenarios to ensure that they can successfully handle those and to then subsequently check whether and how these definitions can be applied to increasingly complex scenarios.

One could even argue that testing causal-model-based definitions of causation or causal explanation on complex, real-life scenarios is not even necessary for corroborating their general adequacy. First, one could invoke the hypothesis that an increasing complexity will not yield entirely new problems. Instead, an increased complexity of a scenario only yields an agglomeration of those problems that we find isolated in simplified toy models. Accordingly, if an explication yields adequate results in all relevant toy models, it will also yield adequate results in more complex real-life scenarios. The hypothesis is not implausible. But whether it is actually true would clearly need some further corroboration.[29] And the obvious way to corroborate it is by testing the causal-model-based definitions on more complex, real-life scenarios after all.

The more promising argument is based on a psychological hypothesis. According to this hypothesis, what we have called 'toy models' are no toy models after all, but instead correspond to how our mind actually represents the world in processes of causal and explanatory reasoning. The human mind may very well employ simplified and manageable causal models of overly complex real-life scenarios to filter only the information that is needed for the identification and evaluation of causes and causal explanations. If this is indeed the case, then causal-model-based definitions do not even need to be applicable to complex, real-life scenarios to be generally adequate explications of causal concepts. These concepts were never supposed to be applied to complex, real-life scenarios in the first place, which would mean that, whenever we talk about actual causes or causal explanations, we do so, not relative to a real-life scenario, but relative to a simplified model of the real-life scenario. But here again, whether this psychological hypothesis is indeed true, still needs empirical corroboration.[30] So, in either way, there is still work to do for substantiating the general validity of causal-model-based accounts of causation and causal explanation, even if the accounts already proved to be successful when applied to rather simple toy examples.

At the end of chapter 8, I have already indicated that empirical psychological research on how human beings actually evaluate the power of explanations could provide valuable data to corroborate, to refute, or to enhance a philosophical explication of explanatory power. The same clearly holds for other explications of causal and explanatory concepts. It is increasingly recognized that empirical research can underpin and enhance philosophical theories of causation and explanation.[31] In the present dissertation, I have already drawn upon empirical studies

state, there is no uncertainty about which structural equations hold in a given situation.

[29] See, for example (Glymour et al., 2010, p. 184 f.), where this hypothesis is doubted.

[30] Work by David Danks already points in this direction. Briefly put, Danks (2014) argues that the human mind actually employs directed acyclic graphs for several distinct cognitive operations and that it is this simplifying representation that enables humans to focus on what is really relevant for the tasks at hand.

[31] See, for example, (Colombo, 2017) for a plea for "an experimental approach to the philosophy of explanation" (Colombo, 2017, p. 514).

in several different subjects.[32] But there is probably and hopefully more empirical results to come. It is my hope, that the philosophical accounts that I have developed in the course of this dissertation may be confronted with further empirical research results and may perhaps even encourage further empirical studies in the field of causation and causal explanation.

Finally, one essential question remains to be answered. How can future work build on the foundation, that I hope to have laid with this dissertation, to advance the project of formalizing IBE? I want to focus on two issues that need to be solved. First, I have only explicated two benchmarks of explanatory goodness: a measure of causal explanatory power and a criterion of proportionality. But I have also acknowledged that there are further dimensions of explanatory goodness, including, for example, the plausibility of the explanatory hypothesis itself.[33] A crucial task that remains to be done, is to stipulate how all the different dimensions of explanatory goodness should be taken into account in order to determine the *best* explanation. As soon as this is done, another crucial question arises, namely whether and how the resulting method of IBE is compatible with other modes of ampliative reasoning. Especially the relationship with Bayesian reasoning has been hotly debated ever since van Fraassen (1989) argued that IBE and Bayesianism are incompatible.[34] Closely related with this issue is the question of whether IBE should actually raise the degree of belief in the selected explanatory hypothesis, as, for example, Niiniluoto (2018) argues, or whether it merely designates the selected hypothesis as being worthy of further pursuit, as, for example, argued by Gabbay and Woods (2005). I am convinced that a formal explication of IBE, that is based on precise formal explications of causal explanation and explanatory power as developed in this dissertation, can ultimately help settle these issues.

[32] For example, (Waskan et al., 2014) and (Wilkenfeld and Lombrozo, 2020) concerning the association of understanding and explanation, (Blanchard et al., forthcoming) concerning causal exclusion, (Walsh and Sloman, 2011) concerning the differentiation of prevention and production of the complement, and of course, (Cheng, 1997), (Liljeholm and Cheng, 2007), and (Buehner et al., 2003) concerning the concept of intrinsic causal power.

[33] See, for example, (Ylikoski and Kuorikoski, 2010) for further dimensions of explanatory goodness.

[34] See, for example, (Douven, 1999), (Weisberg, 2009), (Douven, 2013), (Roche and Sober, 2013), (Douven and Schupbach, 2015), (Climenhaga, 2017), (Cabrera, 2017), (Trpin and Pellert, 2019), which is just a small selection of publications on this issue.

Bibliography

A. Aliseda. *Seeking Explanations: Abduction in Logic, Philosophy of Science and Artificial Intelligence.* ILLC Dissertation Series 1997-4. PhD thesis, Published by the Institute for Logic, Language, and Computation (ILLC), University of Amsterdam, The Netherlands, 1997.

A. Aliseda. Abduction as Epistemic Change: A Peircean Model in Artificial Intelligence. In P. A. Flach and A. C. Kakas, editors, *Abduction and Induction. Applied Logic Series, vol 18.* Dordrecht: Springer, 2000.

H. Andreas and M. Günther. A Ramsey Test Analysis of Causation for Causal Models. *The British Journal for the Philosophy of Science*, 72(2):587–615, 2021.

T. Bartelborth. Explanatory Unification. *Synthese*, 130(1):91–108, 2002.

D. Batens and D. Provijn. Pushing the Search Paths in the Proofs. A Study in Proof Heuristics. *Logique et Analyse*, 173:113–134, 2003.

R. W. Batterman. Multiple Realizability and Universality. *The British Journal for the Philosophy of Science*, 51(1):115–145, 2000.

M. Baumgartner. Interventionism and Epiphenomenalism. *Canadian Journal of Philosophy*, 40 (3):359–383, 2010.

M. Baumgartner. Rendering Interventionism and Non-Reductive Physicalism Compatible. *Dialectica*, 67(1):1–27, 2013.

M. Baumgartner. The Inherent Empirical Underdetermination of Mental Causation. *Australasian Journal of Philosophy*, 96(2):335–350, 2018.

M. Baumgartner and A. Gebharter. Constitutive Relevance, Mutual Manipulability, and Fat-Handedness. *The British Journal for the Philosophy of Science*, 67:731–756, 2016.

S. Beckers. Causal Sufficiency and Actual Causation. *Journal of Philosophical Logic*, 50(6): 1341–1374, 2021.

S. Beckers and J. Vennekens. A Principled Approach to Defining Actual Causation. *Synthese*, 195(2):835–862, 2018.

J. Berkovitz. The Propensity Interpretation of Probability: A Re-evaluation. *Erkenntnis*, 80 (S3):629–711, 2015.

T. Blanchard. Explanatory Abstraction and the Goldilocks Problem: Interventionism Gets Things Just Right. *The British Journal for the Philosophy of Science*, 71(2):633–663, 2020.

T. Blanchard and J. Schaffer. Cause without Default. In H. Beebee, C. Hitchcock, and H. Price, editors, *Making a Difference*, pages 175–214. Oxford: Oxford University Press, 2017.

T. Blanchard, D. Murray, and T. Lombrozo. Experiments on Causal Exclusion. *Mind and Language*, https://doi.org/10.1111/mila.12343, forthcoming.

J. Borner. Halpern and Pearl's Definition of Explanation Amended. *The British Journal for the Philosophy of Science*, https://doi.org/10.1086/716768, forthcoming.

C. Boutilier and V. Becher. Abduction as Belief Revision. *Artificial Intelligence*, 77(1):43–94, 1995.

R. Briggs. The Metaphysics of Chance. *Philosophical Compass*, 5(11):938–952, 2010.

R. Briggs. Interventionist Counterfactuals. *Philosophical Studies*, 160(1):139–166, 2012.

A. Broadbent. *Philosophy of Epidemiology.* London: Palgrave Macmillan, 2013.

G. Brun. Explication as a Method of Conceptual Re-engineering. *Erkenntnis*, 81(6):1211–1241, 2016.

M. J. Buehner, P. W. Cheng, and D. Clifford. From Covariation to Causation: A Test of the Assumption of Causal Power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1119–1140, 2003.

F. Cabrera. Can There Be a Bayesian Explanationism? On the Prospects of a Productive Partnership. *Synthese*, 194(4):1245–1272, 2017.

R. Carnap. *Logical Foundations of Probability.* First edition. Chicago: University of Chicago Press, 1950.

N. Cartwright. Nature's Capacities and Their Measurement. *Oxford: Oxford University Press*, 1989.

U. Chajewska and J. Y. Halpern. Defining Explanation in Probabilistic Systems. *Proceedings of the Thirteenth Conference on Uncertainty in Artifial Intelligence*, pages 62–71, 1997.

P. W. Cheng. From Covariation to Causation: A Causal Power Theory. *Psychological Review*, 104(2):367–405, 1997.

P. W. Cheng. Causality in the Mind: Estimating Contextual and Conjunctive Causal Power. In F. Keil and R. Wilson, editors, *Explanation and Cognition*, pages 227–253. Cambridge, MA: MIT Press, 2000.

P. W. Cheng and L. R. Novick. Assessing Interactive Causal Influence. *Psychological Review*, 111(2):455–485, 2004.

P. W. Cheng and L. R. Novick. Constraints and Nonconstraints in Causal Learning: Reply to White (2005) and to Luhmann and Ahn (2005). *Psychological Review*, 112(3):694–707, 2005.

P. W. Cheng, L. R. Novick, M. Liljeholm, and C. Ford. Explaining Four Psychological Asymmetries in Causal Reasoning: Implications of Causal Assumptions for Coherence. In M. O'Rourke, editor, *Topics in Contemporary Philosophy (Vol. 4): Explanation and Causation*. Cambridge, MA: MIT Press, 2007.

P. W. Cheng, M. Liljeholm, and C. M. Sandhofer. Logical Consistency and Objectivity in Causal Learning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 2034–2039. Austin, Texas: Cognitive Science Society, 2013.

N. Climenhaga. Inference to the Best Explanation Made Incoherent. *Journal of Philosophy*, 114 (5):251–273, 2017.

N. Climenhaga. The Structure of Epistemic Probabilities. *Philosophical Studies*, 177(11):3213–3242, 2020.

M. P. Cohen. On Schupbach and Sprenger's Measures of Explanatory Power. *Philosophy of Science*, 82(1):97–109, 2015.

M. P. Cohen. On Three Measures of Explanatory Power with Axiomatic Representations. *The British Journal for the Philosophy of Science*, 67(4):1077–1089, 2016.

M. P. Cohen. Explanatory Justice: The Case of Disjunctive Explanations. *Philosophy of Science*, 85(3):442–454, 2018.

M. Colombo. Experimental Philosophy of Explanation Rising: The Case for a Plurality of Concepts of Explanation. *Cognitive Science*, 41(2):503–517, 2017.

M. Colombo, M. Postma-Nilsenova, and J. Sprenger. Explanatory Judgment, Probability, and Abductive Inference. In A. Papafragou, D. Grodner, D. Mirman, and J. C. Trueswell, editors, *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pages 432–437. Austin, TX: Cognitive Science Society, 2016.

C. F. Craver. Explaining the Brain. *Oxford: Oxford University Press*, 2007.

V. Crupi. An Argument for Not Equating Confirmation and Explanatory Power. *The Reasoner*, 6(3):39–40, 2012.

V. Crupi and K. Tentori. A Second Look at the Logic of Explanatory Power (with Two Novel Representation Theorems). *Philosophy of Science*, 79(3):365–385, 2012.

D. Danks. *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge, MA: MIT Press, 2014.

H. W. de Regt. *Understanding Scientific Understanding*. Oxford: Oxford University Press, 2017.

H. W. de Regt and D. Dieks. A Contextual Approach to Scientific Understanding. *Synthese*, 144(1):137–170, 2005.

I. Douven. Inference to the Best Explanation Made Coherent. *Philosophy of Science*, 66(S3): 424–435, 1999.

I. Douven. Inference to the Best Explanation, Dutch Books, and Inaccuracy Minimisation. *Philosophical Quarterly*, 63(252):428–444, 2013.

I. Douven and J. N. Schupbach. The Role of Explanatory Considerations in Updating. *Cognition*, 142:299–311, 2015.

P. Dowe. Physical Causation. *Cambridge: Cambridge University Press*, 2000.

P. Dowe. A Counterfactual Theory of Prevention and 'Causation' by Omission. *Australasian Journal of Philosophy*, 79(2):216–226, 2001.

P. Dowe. Causal Processes. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)*. URL = <https://plato.stanford.edu/archives/sum2018/entries/causation-process/>, 2018.

E. Eells. *Probabilistic Causality.* Cambridge, England: Cambridge University Press, 1991.

D. Ehring. Physical Causation. *Mind*, 112(447):529–533, 2003.

N. Emery. Chance, Possibility, and Explanation. *The British Journal for the Philosophy of Science*, 66(1):95–120, 2015.

M. Eronen and D. Brooks. Interventionism and Supervenience: A New Problem and Provisional Solution. *International Studies in the Philosophy of Science*, 28(2):185–202, 2014.

B. Eva and R. Stern. Causal Explanatory Power. *The British Journal for the Philosophy of Science*, 70(4):1029–1050, 2019.

D. Fair. Causation and the Flow of Energy. *Erkenntnis*, 14(3):219–250, 1979.

L. Fenton-Glynn. A Proposed Probabilistic Extension of the Halpern and Pearl Definition of 'Actual Cause'. *The British Journal for the Philosophy of Science*, 68(4):1061–1124, 2017.

B. Fitelson and C. Hitchcock. Probabilistic Measures of Causal Strength. In P. M. Illari, F. Russo, and J. Williamson, editors, *Causality in the Sciences*, pages 600–627. Oxford: Oxford University Press, 2011.

M. Forster, G. Raskutti, R. Stern, and N. Weinberger. The Frugal Inference of Causal Relations. *The British Journal for the Philosophy of Science*, 69(3):821–848, 2018.

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

M. Friedman. Explanation and Scientific Understanding. *Journal of Philosophy*, 71(1):5–19, 1974.

D. M. Gabbay and J. Woods. *The Reach of Abduction: Insight and Trial. A Practical Logic of Cognitive Systems. Volume 2.* Elsevier, 2005.

P. Gärdenfors. Knowledge in Flux: Modeling the Dynamics of Epistemic States. *Cambridge, MA: MIT Press*, 1988.

A. Gebharter. Causal Exclusion and Causal Bayes Nets. *Philosophy and Phenomenological Research*, 95(2):353–375, 2017.

D. H. Glass. Coherence Measures and Inference to the Best Explanation. *Synthese*, 157(3): 275–296, 2007.

S. Glennan. *The New Mechanical Philosophy.* Oxford: Oxford University Press, 2017.

C. Glymour. Learning Causes: Psychological Explanations of Causal Explanation. *Minds and Machines*, 8(1):39–60, 1998.

C. Glymour. Probability and the Explanatory Virtues. *The British Journal for the Philosophy of Science*, 66(3):591–604, 2015.

C. Glymour, D. Danks, B. Glymour, F. Eberhardt, J. Ramsey, R. Scheines, P. Spirtes, C. M. Teng, and J. Zhang. Actual Causation: A Stone Soup Essay. *Synthese*, 175(2):169–192, 2010.

L. Glynn. Deterministic Chance. *The British Journal for the Philosophy of Science*, 61(1):51–80, 2010.

I. J. Good. Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments. *Journal of the Royal Statistical Society. Series B*, 22(2):319–331, 1960.

I. J. Good. A Causal Calculus I. *The British Journal for the Philosophy of Science*, 11(44): 305–318, 1961a.

I. J. Good. A Causal Calculus II. *The British Journal for the Philosophy of Science*, 12(45): 43–51, 1961b.

A. Hájek. What Conditional Probability Could Not Be. *Synthese*, 137(3):273–323, 2003.

A. Hájek. Interpretations of Probability. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Fall 2019 Edition).* URL = <https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/>, 2019.

N. Hall. Two Concepts of Causation. In J. Collins, N. Hall, and L. A. Paul, editors, *Causation and Counterfactuals*, pages 225–276. Cambridge, MA: MIT Press, 2004.

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

J. Y. Halpern. A Modification of the Halpern-Pearl Definition of Causality. *Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 3022–3033, 2015.

J. Y. Halpern. Actual Causality. *Cambridge, MA: MIT Press*, 2016.

J. Y. Halpern and C. Hitchcock. Actual Causation and the Art of Modeling. In R. Dechter, H. Geffner, and J. Y. Halpern, editors, *Causality, Probability, and Heuristics: A Tribute to Judea Pearl*, pages 383–406. London: College Publications, 2010.

J. Y. Halpern and C. Hitchcock. Graded Causation and Defaults. *The British Journal for the Philosophy of Science*, 66(2):413–457, 2015.

J. Y. Halpern and J. Pearl. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, 56(4):843–887, 2005a.

J. Y. Halpern and J. Pearl. Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005b.

T. Handfield, C. R. Twardy, K. B. Korb, and G. Oppy. The Metaphysics of Causal Models: Where's the Biff? *Erkenntnis*, 68(2):149–168, 2008.

G. H. Harman. The Inference to the Best Explanation. *Philosophical Review*, 74(1):88–95, 1965.

D. M. Hausmann. Physical causation. *Studies in History and Philosophy of Modern Physics*, 33 (B):717–724, 2002.

C. G. Hempel. Aspects of Scientific Explanation. In C. G. Hempel, editor, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press, 1965.

C. G. Hempel and P. Oppenheim. Studies in the Logic of Explanation. *Philosophy of Science*, 15(2):135–175, 1948.

G. Hesslow. Two Notes on the Probabilistic Approach to Causality. *Philosophy of Science*, 43 (2):290–292, 1976.

E. Hiddleston. Causal Powers. *The British Journal for the Philosophy of Science*, 56(1):27–59, 2005.

J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Ithaca: Cornell University Press, 1962.

C. Hitchcock. Salmon on Explanatory Relevance. *Philosophy of Science*, 62(2):304–320, 1995.

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

C. Hitchcock. Causal Modelling. In H. Beebee, C. Hitchcock, and P. Menzies, editors, *The Oxford Handbook of Causation*. Oxford: Oxford University Press, 2009a.

C. Hitchcock. Problems for the Conserved Quantity Theory: Counterexamples, Circularity, and Redundancy. *The Monist*, 92(1):72–93, 2009b.

C. Hitchcock. Theories of Causation and the Causal Exclusion Argument. *Journal of Consciousness Studies*, 19(5):40–56, 2012.

C. Hitchcock and J. Woodward. Explanatory Generalizations, Part II: Plumbing Explanatory Depth. *Noûs*, 37(2):181–199, 2003.

C. Hoefer. The Third Way on Objective Probability: A Skeptic's Guide to Objective Chance. *Mind*, 116(2):549–596, 2007.

P. Humphreys. The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences. *Princeton: Princeton University Press*, 1989.

A. Hüttemann. The Return of Causal Powers? In S. Psillos, H. Lagerlund, and B. Hill, editors, *Causal Powers in Science: Blending Historical and Conceptual Perspectives*, pages 168–185. Oxford University Press, 2021.

P. Illari. Why Theories of Causality Need Production: An Information Transmission Account. *Philosophy and Technology*, 24(2):95–114, 2011.

P. Illari. Mechanistic Explanation: Integrating the Ontic and Epistemic. *Erkenntnis*, 78(2): 237–255, 2013.

J. Kim. *Physicalism, or Something Near Enough.* Princeton: Princeton University Press, 2005.

D. Kinney. On the Explanatory Depth and Pragmatic Value of Coarse-Grained, Probabilistic, Causal Explanations. *Philosophy of Science*, 86(1):145–167, 2019a.

D. Kinney. *The Problem of Granularity for Scientific Explanation.* PhD thesis, London School of Economics and Political Science, URL = <http://etheses.lse.ac.uk/id/eprint/3996>, 2019b.

P. Kitcher. Explanation, Conjunction, and Unification. *Journal of Philosophy*, 73(8):207–212, 1976.

P. Kitcher. Explanatory Unification. *Philosophy of Science*, 48(4):507 – 531, 1981.

P. Kitcher. Explanatory Unification and the Causal Structure of the World. In P. Kitcher and W. C. Salmon, editors, *Scientific Explanation*, pages 410–505. Minneapolis: University of Minnesota Press, 1989.

H. Klärner. *Der Schluß auf die beste Erklärung.* Berlin: De Gruyter, 2003.

M. Lange. What Makes a Scientific Explanation Distinctively Mathematical? *The British Journal for the Philosophy of Science*, 64(3):485–511, 2013.

H. Leitgeb. A Probabilistic Semantics for Counterfactuals. Part A. *Review of Symbolic Logic*, 5 (1):85–121, 2012.

H. Leitgeb and A. Carus. Rudolf Carnap. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Fall 2022 Edition).* URL = <https://plato.stanford.edu/archives/fall2022/entries/carnap/>, 2022.

D. Lewis. Postscripts to 'Causation'. In *Philosophical Papers, Volume II*, pages 173–213. Oxford: Oxford University Press, 1986a.

D. Lewis. A Subjectivist's Guide to Objective Chance. In *Philosophical Papers, Volume II.* Oxford: Oxford University Press, 1986b.

D. Lewis. Causal Explanation. In *Philosophical Papers, Volume II.* Oxford: Oxford University Press, 1986c.

D. Lewis. Humean Supervenience Debugged. *Mind*, 103(412):473–490, 1994.

D. Lewis. Void and Object. In J. Collins, N. Hall, and L. A. Paul, editors, *Causation and Counterfactuals*, pages 277–290. Cambridge, MA: MIT Press, 2004.

M. Liljeholm and P. W. Cheng. When Is a Cause the 'Same'? Coherent Generalization across Contexts. *Psychological Science*, 18(11):1014–1021, 2007.

P. Lipton. *Inference to the Best Explanation*. London: Routledge, second edition, 2004.

C. List. Levels: Descriptive, Explanatory, and Ontological. *Noûs*, 53(4):852–883, 2019.

C. List and P. Menzies. Nonreductive Physicalism and the Limits of the Exclusion Principle. *Journal of Philosophy*, 106(9):475–502, 2009.

C. List and M. Pivato. Emergent Chance. *Philosophical Review*, 124(1):119–152, 2015.

P. K. Machamer, L. Darden, and C. F. Craver. Thinking About Mechanisms. *Philosophy of Science*, 67(1):1–25, 2000.

J. L. Mackie. *The Cement of the Universe*. Oxford: Oxford University Press, 1974.

T. Maudlin. Causation, Counterfactuals, and the Third Factor. In J. Collins, N. Hall, and L. A. Paul, editors, *Causation and Counterfactuals*, pages 419–443. MIT Press, 2004.

T. McGrew. Confirmation, Heuristics, and Explanatory Reasoning. *The British Journal for the Philosophy of Science*, 54(4):553–567, 2003.

J. Meheus. A Formal Logic for the Abduction of Singular Hypotheses. In D. Dieks, W. Gonzalez, S. Hartmann, T. Uebel, and M. Weber, editors, *Explanation, Prediction, and Confirmation. The Philosophy of Science in a European Perspective, vol. 2*, pages 93–108. Dordrecht: Springer, 2011.

J. Meheus and D. Batens. A Formal Logic for Abductive Reasoning. *Logic Journal of the IGPL*, 14(2):221–236, 2006.

J. Meheus and D. Provijn. Abduction through Semantic Tableaux versus Abduction through Goal-Directed Proofs. *Theoria*, 22(3):295–304, 2007.

P. Menzies. The Problem of Counterfactual Isomorphs. In H. Beebee, C. Hitchcock, and H. Price, editors, *Making a Difference: Essays on the Philosophy of Causation*, pages 153–174. Oxford: Oxford University Press, 2017.

R. Neapolitan and X. Jiang. The Bayesian Network Story. In A. Hájek and C. Hitchcock, editors, *The Oxford Handbook of Probability and Philosophy*, pages 183–200. Oxford: Oxford University Press, 2017.

I. Niiniluoto. *Truth-Seeking by Abduction*. Cham, Switzerland: Springer, 2018.

M. Pagnucco. *The Role of Abductive Reasoning within the Process of Belief Revision.* PhD thesis, Basser Department of Computer Science, University of Sidney, Australia, URL=<http://www.cse.unsw.edu.au/ morri/Papers/morri.PhD.pdf>, 1996.

J. Pearl. Probabilities of Causation: Three Counterfactual Interpretations and Their Identification. *Synthese*, 121(1-2):93–149, 1999.

J. Pearl. Causality: Models, Reasoning and Inference. *Cambridge: Cambridge University Press*, 2000.

T. W. Polger, L. A. Shapiro, and R. Stern. In Defense of Interventionist Solutions to Exclusion. *Studies in History and Philosophy of Science. Part A*, 68:51–57, 2018.

K. Popper. *The Logic of Scientific Discovery.* London, New York: Routledge, 1959.

A. Reutlinger. Is There A Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation. *Philosophy of Science*, 83(5):733–745, 2016.

A. Reutlinger. Does the Counterfactual Theory of Explanation Apply to Non-Causal Explanations in Metaphysics? *European Journal for Philosophy of Science*, 7(2):239–256, 2017.

A. Reutlinger, M. Colyvan, and K. Krzyzanowska. The Prospects for a Monist Theory of Non-Causal Explanation in Science and Mathematics. *Erkenntnis*, 87(4):1773–1793, 2020.

W. Roche and E. Sober. Explanatoriness is Evidentially Irrelevant, or Inference to the Best Explanation Meets Bayesian Confirmation Theory. *Analysis*, 73(4):659–668, 2013.

W. C. Salmon. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press, 1984.

W. C. Salmon. *Four Decades of Scientific Explanation.* Pittsburgh: University of Pittsburgh Press, 1989.

W. C. Salmon. Causality and Explanation: A Reply to Two Critiques. *Philosophy of Science*, 64(3):461–477, 1997.

J. Schaffer. Causes Need Not be Physically Connected to Their Effects: The Case for Negative Causation. In C. Hitchcock, editor, *Contemporary Debates in Philosophy of Science*, pages 197–216. Oxford: Blackwell, 2004.

J. Schaffer. Grounding in the Image of Causation. *Philosophical Studies*, 173(1):49–100, 2016.

J. Schaffer. Laws for Metaphysical Explanation. *Philosophical Issues*, 27(1):302–321, 2017.

J. N. Schupbach. Comparing Probabilistic Measures of Explanatory Power. *Philosophy of Science*, 78(5):813–829, 2011a.

J. N. Schupbach. *Studies in the Logic of Explanatory Power.* PhD thesis, University of Pittsburgh, URL=<http://d-scholarship.pitt.edu/id/eprint/7885>, 2011b.

J. N. Schupbach and J. Sprenger. The Logic of Explanatory Power. *Philosophy of Science*, 78 (1):105–127, 2011.

G. Schurz and K. Lambert. Outline of a Theory of Scientific Understanding. *Synthese*, 101(1): 65–120, 1994.

W. Schwarz. Best System Approaches to Chance. In A. Hájek and C. Hitchcock, editors, *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press, 2016.

L. A. Shapiro and E. Sober. Against Proportionality. *Analysis*, 72(1):89–93, 2012.

P. Spirtes and R. Scheines. Causal Inference of Ambiguous Manipulations. *Philosophy of Science*, 71(5):833–845, 2004.

P. Spirtes, C. Glymour, and R. Scheines. Causation, Prediction, and Search. *Cambridge, MA: MIT Press*, 2000.

J. Sprenger. Foundations of a Probabilistic Theory of Causal Strength. *Philosophical Review*, 127(3):371–398, 2018.

J. Sprenger and S. Hartmann. Bayesian Philosophy of Science. *Oxford: Oxford University Press*, 2019.

J. Sprenger and J. Stegenga. Three Arguments for Absolute Outcome Measures. *Philosophy of Science*, 84(5):840–852, 2017.

J. Stegenga. Measuring Effectiveness. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 54:62–71, 2015.

M. Steiner. Mathematical Explanation. *Philosophical Studies*, 34(2):135–151, 1978.

S. Stephan and M. R. Waldmann. Preemption in Singular Causation Judgments: A Computational Model. *Topics in Cognitive Science*, 10(1):242–257, 2018.

S. Stephan and M. R. Waldmann. The Role of Mechanism Knowledge in Singular Causation Judgments. *Cognition*, 218:104924, 2022.

S. Stephan, R. Mayrhofer, and M. R. Waldmann. Assessing Singular Causation: The Role of Causal Latencies. In T. Rogers, M. Rau, X. Zhu, and C. Kalish, editors, *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 1080–1085. Austin TX: Cognitive Science Society, 2018.

S. Stephan, R. Mayrhofer, and M. R. Waldmann. Time and Singular Causation - A Computational Model. *Cognitive Science*, 44(7):e12871, 2020.

C. Straßer. *Adaptive Logics for Defeasible Reasoning. Applications in Argumentation, Normative Reasoning and Default Reasoning.* Cham, Switzerland: Springer, 2014.

M. Strevens. *Depth. An Account of Scientific Explanation.* Cambridge, MA: Harvard University Press, 2008.

P. Suppes. *A Probabilistic Theory of Causality.* Amsterdam: North-Holland Publishing Company, 1970.

P. Thagard. Explanatory Coherence. *Behavioral and Brain Sciences*, 12(3):435–467, 1989.

The Bible. *Authorized King James version.* Oxford: Oxford University Press, 1998.

B. Trpin and M. Pellert. Inference to the Best Explanation in Uncertain Evidential Situations. *The British Journal for the Philosophy of Science*, 70(4):977–1001, 2019.

B. C. van Fraassen. *The Scientific Image.* Oxford: Oxford University Press, 1980.

B. C. van Fraassen. *Laws and Symmetry.* Oxford: Oxford University Press, 1989.

N. Vasilyeva, T. Blanchard, and T. Lombrozo. Stable Causal Relationships Are Better Causal Relationships. *Cognitive Science*, 42(4):1265–1296, 2018.

C. R. Walsh and S. A. Sloman. The Meaning of Cause and Prevent: The Role of Causal Mechanism. *Mind and Language*, 26(1):21–52, 2011.

J. Waskan, I. Harmon, Z. Horne, J. Spino, and J. Clevenger. Explanatory Anti-Psychologism Overturned by Lay and Scientific Case Classifications. *Synthese*, 191(5):1013–1035, 2014.

J. Weisberg. Locating IBE in the Bayesian Framework. *Synthese*, 167(1):125–143, 2009.

B. Weslake. Exclusion Excluded. *Manuscript. URL=<https://philpapers.org/rec/wesee>*, 2011.

B. Weslake. Proportionality, Contrast and Explanation. *Australasian Journal of Philosophy*, 91 (4):785–797, 2013.

B. Weslake. A Partial Theory of Actual Causation. *The British Journal for the Philosophy of Science*, Preprint retrieved from: URL=<http://philpapers.org/rec/wesapt>, forthcoming.

D. A. Wilkenfeld and T. Lombrozo. Explanation Classification Depends on Understanding: Extending the Epistemic Side-Effect Effect. *Synthese*, 197(6):2565–2592, 2020.

J. Williamson. Bayesian Nets and Causality: Philosophical and Computational Foundations. *Oxford: Oxford University Press*, 2004.

J. Wilson. Determinables and Determinates. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Spring 2021 Edition).* URL = <https://plato.stanford.edu/archives/spr2021/entries/determinate-determinables/>, 2021.

J. Woodward. *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press, 2003.

J. Woodward. Sensitive and Insensitive Causation. *The Philosophical Review*, 115(1):1–50, 2006.

J. Woodward. Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation. *Biology and Philosophy*, 25(3):287–318, 2010.

J. Woodward. 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.

J. Woodward. Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, 91(2):303–347, 2015.

J. Woodward. The Problem of Variable Choice. *Synthese*, 193(4):1047–1072, 2016.

J. Woodward. Causal Cognition: Physical Connections, Proportionality, and the Role of Normative Theory. In W. J. Gonzalez, editor, *Philosophy of Psychology: Causality and Psychological Subject*, pages 105–138. Berlin, Boston: De Gruyter, 2018.

J. Woodward. Explanatory Autonomy: The Role of Proportionality, Stability, and Conditional Irrelevance. *Synthese*, 198(1):237–265, 2021.

J. Woodward and C. Hitchcock. Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs*, 37(1):1–24, 2003.

J. Woodward and L. Ross. Scientific Explanation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*. URL = <https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/>, 2021.

S. Yablo. Mental Causation. *Philosophical Review*, 101(2):245–280, 1992.

S. Yablo. De Facto Dependence. *The Journal of Philosophy*, 99(3):130–148, 2002.

P. Ylikoski and J. Kuorikoski. Dissecting Explanatory Power. *Philosophical Studies*, 148(2): 201–209, 2010.

Zusammenfassung

Innerhalb der Wissenschaftstheorie, aber auch der analytisch geprägten Philosophie im Allgemeinen, erhält die Ansicht immer mehr Anerkennung, dass die Methode des Schlusses auf die beste Erklärung ein wesentliches Werkzeug unseres Denkvermögens ist, das sowohl in banalen Alltagssituationen, als auch in komplexen und folgenschweren Entscheidungslagen zur Anwendung kommt.[35] Im Laufe der letzten 25 Jahre kam es daher auch vermehrt zu Versuchen, die Methode des Schlusses auf die beste Erklärung formal genau zu explizieren. Boutilier und Becher (1995) sowie Pagnucco (1996) versuchen beispielsweise Schlüsse auf die beste Erklärung anhand der AGM-Theorie der Glaubensrevision formal zu erfassen.[36] Aliseda (1997) sowie Meheus und Provijn (2007) formulieren, jeweils auf Grundlage eines anderen formalen Beweisverfahrens, einen Algorithmus, der für ein gegebenes zu erklärendes Ereignis (Explanandum) und eine gegebene Hintergrundtheorie verschiedene Erklärungskandidaten an die Hand gibt. Aufbauend auf dem formalen System der adaptiven Logik schlagen Meheus und Batens (2006) noch eine weitere Formalisierung von Schlüssen auf die beste Erklärung vor, die etwas später von Meheus (2011) weiterentwickelt wird.[37]

All diese Ansätze Schlüsse auf die beste Erklärung formal zu explizieren, teilen allerdings eine sehr bedeutende Schwachstelle. Sie alle setzen die Grundannahmen eines Erklärungsmodells voraus, das innerhalb der philosophischen Literatur längst als überholt gilt. Die Rede ist vom deduktiv-nomologischen (DN) Erklärungsmodell von Carl Gustav Hempel und Paul Oppenheim (1948), das im Wesentlichen besagt, dass eine Hypothese ein gegebenes Explanandum erklärt, wenn sich das Explanandum von der Hypothese und gewissen weiteren Hintergrundannahmen durch ein deduktives Argument ableiten lässt.[38] Innerhalb der letzten Jahrzehnte hat die philosophische Literatur jede Menge Beispiele hervorgebracht, in denen das DN-Modell etwas als Erklärung identifiziert, was intuitiv eindeutig keine Erklärung ist. Entsprechend problematisch sind die oben genannten Formalisierungen von Schlüssen auf die beste Erklärung. Keiner der Ansätze stellt zuverlässig sicher, dass die Hypothesen, auf die geschlossen wird, intuitiv auch wirklich Erklärungen des gegebenen Explanandums sind.

In meiner Dissertation werde ich keine vollständige Formalisierung von Schlüssen auf die beste Erklärung liefern. Aufgrund der dafür noch ausstehenden Vorarbeit, wäre dies ein viel zu umfassendes Projekt um es im Rahmen einer einzigen Dissertation zu vollenden. Stattdessen ist es mein Ziel, mit meiner Dissertation ein solides Fundament zu schaffen, das für eine formale

[35] Siehe beispielsweise (Harman, 1965), (Lipton, 2004), (Gabbay and Woods, 2005), (Niiniluoto, 2018), or (Douven and Schupbach, 2015).

[36] Siehe (Gärdenfors, 1988) für eine ausführliche Darstellung der AGM-Theorie der Glaubensrevision.

[37] Siehe (Straßer, 2014) für eine umfangreiche Einführung in die adaptive Logik.

[38] Siehe auch (Hempel, 1965) für eine ausführliche Darstellung des DN-Modells.

Explikation von Schlüssen auf die beste Erklärung nötig ist. Dieses Fundament besteht letztlich aus drei Teilen: Erstens, aus einem formal präzisen, zweckmäßigen und intuitiv adäquaten Erklärungsmodell. Zweitens, aus einem formal präzisen, zweckmäßigen und intuitiv adäquaten Maßstab der Erklärungsgüte. Und drittens, aus einem formal präzisen, zweckmäßigen und intuitiv adäquaten Kriterium der Proportionalität einer Erklärung.

Teil I: Ein formales Erklärungsmodell

Eine wesentliche Motivation für meine Arbeit basiert auf der Überzeugung, dass eine angemessene formale Modellierung von Schlüssen auf die beste Erklärung zunächst ein angemessenes formales Modell von Erklärungen braucht. Es ist das erste zentrale Ziel dieser Dissertation, ein solches Modell vorzulegen.

Nachdem das DN-Modell seinen Status als fest etabliertes Erklärungsmodell verloren hatte, ist eine über mehrere Jahrzehnte andauernde Debatte darüber ausgebrochen, wie ein angemessenes Erklärungsmodell aussehen sollte.[39] Auch wenn sich seither kein Modell mehr als allein vorherrschendes Erklärungsmodell durchsetzen konnte, hat sich doch so etwas wie ein neuer Konsens gebildet. Dieser neue Konsens besagt, dass kausale Erklärungsmodelle, also Modelle, denen zufolge eine Erklärung in der Angabe bestimmter Ursachen des Explanandums besteht, dazu in der Lage sind, die größten Probleme des DN-Modells zufriedenstellend zu lösen.[40] In meiner Dissertation setze ich nun an einem konkreten kausalen Erklärungsmodell an, das in der philosophischen Literatur bislang kaum Beachtung gefunden hat, nämlich an dem Erklärungsmodell von Halpern und Pearl (2005b), das von Halpern (2016) noch einmal leicht überarbeitet wurde. Es gibt vor allem drei Gründe, warum das Modell von Halpern und Pearl (HP-Modell) den Ausgangspunkt meiner Arbeit bildet.

Erstens: Halpern und Pearl bauen ihr Erklärungsmodell auf einem der am besten etablierten und technisch ausgefeiltesten formalen Explikationen von Kausalität auf: Einer interventionistischen Kausaltheorie innerhalb von Strukturgleichungsmodellen.[41] Kausalbeziehungen zwischen Ereignissen werden innerhalb einer solchen Theorie ganz wesentlich anhand von bestimmten kontrafaktischen Konditionalen definiert, deren Wahrheitsbedingungen anhand von Operationen auf Strukturgleichungsmodellen genau definiert werden können.

Zweitens: Das HP-Modell expliziert in erster Linie den Begriff der *potentiellen* Erklärung von konkreten Ereignissen. Es ist also darauf spezialisiert, für ein gegebenes Explanandum solche Ereignisse zu identifizieren, die das Potential haben, das Explanandum zu erklären. Das schließt insbesondere auch solche Ereignisse ein, bei denen noch gar nicht sicher ist, ob sie tatsächlich vorgefallen sind und ob sie das Explanandum daher in der gegebenen Situation auch tatsächlich erklären. Es ist gerade ein solcher Begriff der *potentiellen* Erklärung, der für eine Formalisierung von Schlüssen auf die beste Erklärung gebraucht wird. Denn in einem Schluss auf die beste Erklärung ist natürlich noch nicht bekannt, welche der gegebenen Erklärungskandidaten

[39] Siehe beispielsweise (Salmon, 1989), (Klärner, 2003) oder (Woodward and Ross, 2021) für ausführliche und systematische Darstellungen dieser Debatten.

[40] Gemeint sind die berühmten Irrelevanz- und Asymmetrie-Gegenbeispiele. Siehe beispielsweise (Salmon, 1989, S. 46 ff.).

[41] Siehe (Pearl, 2000), (Woodward, 2003), (Halpern, 2016) für eine detaillierte Darstellung von interventionistischen Kausaltheorien innerhalb von Strukturgleichungsmodellen.

tatsächlich der Fall sind. Andernfalls wäre ein Schluss auf die beste Erklärung ja überhaupt nicht nötig.

Drittens: Das HP-Modell definiert kausale Erklärungen immer relativ zu einem gegebenen epistemischen Zustand. Jede kausale Erklärung ist demnach also immer eine kausale Erklärung *für* einen ganz bestimmten Akteur in einem bestimmten epistemischen Zustand. Diese Berücksichtigung eines epistemischen Zustands ist ebenfalls wesentlich bei Schlüssen auf die beste Erklärung, denn jeder Schluss auf die beste Erklärung für ein gegebenes Phänomen geschieht immer in Anbetracht von ganz bestimmten Hintergrundüberzeugungen. Es sind diese Hintergrundüberzeugungen, die ganz wesentlich mitbestimmen, welche Hypothesen überhaupt als potentielle Erklärungen in Betracht gezogen werden und wie diese Hypothesen als potentielle Erklärungen bewertet werden. Die Relativierung einer Erklärung auf einen gegebenen epistemischen Zustand unterscheidet das HP-Modell von anderen, innerhalb der Philosophie zum Teil deutlich bekannteren, kausalen Erklärungsmodellen, wie dem von Salmon (1984), Lewis (1986c), Woodward (2003), Craver (2007) oder Strevens (2008).

Die drei genannten Aspekte machen das kausale Erklärungsmodell von Halpern und Pearl zu einer vielversprechenden Grundlage für eine Formalisierung von Schlüssen auf die beste Erklärung. Dennoch zeige ich im ersten Kapitel meiner Dissertation, nach einer kurzen Einführung in den formalen Rahmen der Strukturgleichungsmodelle, dass das HP-Modell trotz seiner herausragenden Eigenschaften noch mit einigen Unzulänglichkeiten zu kämpfen hat. Auf Grundlage dieser Kritik erarbeite ich ein neues Erklärungsmodell, das einerseits die genannten positiven Eigenschaften des HP-Modells beibehält, andererseits aber dessen aufgezeigte Unzulänglichkeiten überwindet. Das von mir vorgeschlagene Erklärungsmodell besteht dabei aus gleich mehreren Definitionen, die mehrere miteinander verwandte Erklärungsbegriffe explizieren. Dazu gehören unter anderem die Begriffe der *potentiellen Erklärung*, der *tatsächlichen Erklärung*, der *partiellen Erklärung* und der *korrekten Erklärung*.

Wie auch schon Halpern und Pearl (2005b) und Halpern (2016), habe ich im ersten Kapitel lediglich Kontexte mit deterministischen Kausalbeziehungen betrachtet. Tatsächlich können wir viele Kausalbeziehungen allerdings nur probabilistisch beschreiben. Im zweiten Kapitel widme ich mich deshalb zunächst der Frage, wie in Anlehnung an deterministische Strukturgleichungsmodelle probabilistische Kausalbeziehungen formal modelliert werden können. Hierbei vergleiche ich zwei verwandte, aber verschiedene, formale Ansätze. Der erste Ansatz nutzt Bayes-Netze für die Modellierung von probabilistischen Kausalbeziehungen.[42]. Der zweite Ansatz nutzt hingegen weiterhin deterministische Strukturgleichungsmodelle, führt in diese Modelle aber zusätzliche exogene Variablen ein, um Störfaktoren zu repräsentieren.[43] Ich zeige ausführlich an verschiedenen Beispielen, dass beide formale Ansätze letztlich zu zwei verschiedenen Interpretationen von Kausalität zwischen Einzelereignissen führen. Beide Interpretationen unterscheiden sich nicht nur intensional, sondern auch extensional. Es gibt daher Beispiele von Einzelereignissen, in denen laut einer der beiden Interpretationen eine Kausalbeziehung gesichert vorliegt, während das laut der anderen Interpretation nicht der Fall ist.

Für die Explikation meines Erklärungsmodells nutze ich im weiteren Verlauf ausschließlich

[42] Siehe beispielsweise (Williamson, 2004) für eine umfassende Einführung in Bayes-Netze.

[43] (Pearl, 2000) liefert eine umfassende Beschreibung von Strukturgleichungsmodellen mit Störfaktoren.

Strukturgleichungsmodelle mit Störfaktoren und keine Bayes-Netze. Diese Entscheidung ist darin begründet, dass die Interpretation von Kausalität zwischen Einzelereignissen, wie sie aus Strukturgleichungsmodellen mit Störfaktoren hervorgeht, letztlich besser mit einer psychologischen Theorie über das menschliche Erkennen und Bewerten von Ursachen zusammenpasst, die im weiteren Verlauf meiner Dissertation noch eine bedeutende Rolle spielen wird. In Kapitel 3 zeige ich deshalb innerhalb des formalen Rahmens der Strukturgleichungsmodelle mit Störfaktoren, wie das Erklärungsmodell, das ich im ersten Kapitel entwickelt habe, auch in Kontexten mit probabilistischen Kausalbeziehungen angewandt werden kann.

Mit Abschluss von Kapitel 3 ist das erste wesentliche Ziel meiner Dissertation erreicht: Wir haben nun eine zweckmäßige, intuitiv adäquate, formale Explikation von potentiellen kausalen Erklärungen. Für eine Formalisierung von Schlüssen auf die beste Erklärung ist das aber natürlich noch nicht genug. Schließlich geht es nicht darum, auf *irgendeine* potentielle kausale Erklärung zu schließen. Stattdessen soll auf *die beste* aller vorhandenen potentiellen Erklärungen geschlossen werden. Um diese beste Erklärung ausfindig zu machen, brauchen wir allerdings einen Maßstab, der es uns erlaubt, verschiedene potentielle Erklärungen miteinander zu vergleichen. Es ist deshalb das zweite wesentliche Ziel meiner Dissertation, einen solchen Maßstab zu explizieren. In den Kapiteln 4 bis 8 arbeite ich auf dieses Ziel hin.

Teil II: Erklärungsstärke messen

Da das Erklärungsmodell, das ich im ersten Teil der Dissertation entwickelt habe, eine Explikation von kausalen Erklärungen ist, in denen ein Ereignis durch die Angabe ganz bestimmter Ursachen erklärt wird, ist ein Maßstab, der die Stärke einer Kausalbeziehung misst, ein naheliegender Ausgangspunkt für die Suche nach einem Maßstab, der die Stärke einer kausalen Erklärung bestimmen kann. Im Rahmen ihrer *Power PC* Theorie hat die Psychologin Patricia Cheng (1997) einen solchen Maßstab der *causal power* hergeleitet. Sowohl rein mathematische Überlegungen als auch empirische Studien legen nahe, dass dieser Maßstab dazu in der Lage ist, das zu messen, was Menschen intuitiv als die intrinsische kausale Kraft (*intrinsic causal power*) einer produktiven Ursache (*generative cause*) in Bezug auf einen gegebenen Effekt empfinden.[44] Diese intrinsische kausale Kraft ist deshalb so interessant, weil sie automatisch als invariant gegenüber kontextuellen Veränderungen angenommen wird, was bedeutet, dass die Quantität der Kraft konstant bleibt, auch wenn sich der Kontext, in dem die Ursache vorkommt, verändert. Um quantitative Aussagen über die Konsequenzen einer Ursache in neuen Kontexten zu treffen, ist ein Maßstab, der die intrinsische kausale Kraft dieser Ursache misst, daher von ganz zentraler Bedeutung.

Cheng's Theorie hat allerdings ein Problem. Die Formel, die laut Cheng's Herleitung die intrinsische kausale Kraft einer Ursache messen kann, ist nur in sehr wenigen und recht einfachen kausalen Szenarien anwendbar. In Kapitel 4 erweitere ich deshalb Cheng's *Power PC* Theorie, um einen Maßstab der intrinsischen kausalen Kraft zu entwickeln, der auch in komplexeren Szenarien anwendbar ist. Auf Grundlage dieser Verallgemeinerung zeige ich, dass kürzlich vorgebrachte Kritik an Cheng's Maßstab auf falschen Grundannahmen beruht.

Die intrinsische kausale Kraft einer produktiven Ursache (*generative cause*) hat in Cheng's

[44] Siehe (Fitelson and Hitchcock, 2011), (Cheng, 1997), (Buehner et al., 2003) und (Liljeholm and Cheng, 2007).

Power PC Theorie noch einen Gegenspieler: Die intrinsische kausale Kraft einer verhindernden Ursache (*preventive cause*). Auch hier gilt, dass Cheng's hergeleitete Formel zur Messung der kausalen Kraft einer verhindernden Ursache in nur wenigen und sehr einfachen kausalen Szenarien anwendbar ist. Daher erweitere ich in Kapitel 5 erneut Cheng's *Power PC* Theorie, um einen Maßstab der intrinsischen kausalen Kraft einer verhindernden Ursache zu entwickeln, der auch in komplexeren Szenarien anwendbar ist.

In Cheng's *Power PC* Theorie, und daher auch in unserer erweiterten Variante dieser Theorie, spielen zwei Kausalbegriffe eine ganz wesentliche Rolle: Der Begriff der kausalen Produktion (*causal production*) und der Begriff der kausalen Verhinderung (*prevention*). In Kapitel 6 untersuche ich, wie diese beiden Begriffe mit den interventionistischen Kausalbegriffen zusammenhängen, welche die Grundlage meines Erklärungsmodells aus Kapitel 1 und 3 bilden.

Die Untersuchung aus Kapitel 6 motiviert schließlich die Definition eines weiteren Erklärungsbegriffs, der mein Erklärungsmodell aus Kapitel 1 und 3 ergänzt: Den Begriff der ausführlichen kausalen Erklärung (*extensive causal explanation*). Die Idee, die diesem Begriff zu Grunde liegt, ist die, dass eine ausführliche kausale Erklärung explizit alle Informationen enthält, die für die Bewertung der Stärke der gegebenen Erklärung notwendig sind. Ich entwickle zudem einen Algorithmus, der in einem gegebenen Strukturgleichungsmodell alle ausführlichen kausalen Erklärung eines gegebenen Explanandums ϕ im folgenden Format produziert: ϕ, *weil* ψ, *obwohl* χ.

In Kapitel 8 beschäftige ich mich schließlich mit drei bekannten Bayesschen Maßstäben der Erklärungskraft: Mit dem Maßstab von Schupbach und Sprenger (2011), dem Maßstab von Crupi und Tentori (2012), und dem Maßstab von Good (1960) und McGrew (2003). Ich führe mehrere Argumente an, die darlegen, dass keiner der drei Maßstäbe dazu in der Lage ist, das zu messen, was wir intuitiv unter der Stärke einer Erklärung (*explanatory power*) verstehen.[45] Aufbauend auf den verallgemeinerten Maßstäben für die intrinsische kausale Kraft einer produktiven Ursache und die intrinsische kausale Kraft einer verhindernden Ursache schlage ich deshalb eine neue Methode vor, um die (intrinsische) Kraft einer kausalen Erklärung zu bestimmen. Abschließend zeige ich, dass die Kritikpunkte, die auf die drei genannten Bayesschen Maßstäbe zutreffen, nicht auf meine neue Messmethode der Erklärungskraft zutreffen.

Mit dem Abschluss von Kapitel 8 ist daher auch das zweite Hauptziel meiner Dissertation erreicht: Wir haben nun eine zweckmäßige, intuitiv adäquate, formale Explikation eines Maßstabs, der die Kraft einer kausalen Erklärung messen kann.

Teil II: Die Suche nach dem richtigen Level

Yablo (1992) hat mit dem Begriff der Proportionalität (*proportionality*) ein weiteres intuitiv überzeugendes Kriterium für die Güte einer kausalen Erklärung ins Feld geführt. Das Kriterium der Proportionalität soll sicher stellen, dass eine Ursache für ein gegebenes Explanandum weder zu abstrakt, noch zu spezifisch beschrieben wird. Allerdings ist es innerhalb eines klassischen Strukturgleichungsmodells gar nicht möglich, kausale Beziehungen auf verschiedenen Ebenen der Abstraktheit, also auf verschiedenen Ebenen der Supervenienz zu formulieren. Unter Einbeziehung eines Formalismus, der von List (2019) entwickelt wurde, schlage ich in Kapitel 9 de-

[45] Ein Teil dieser Argumente wurde bereits von Eva und Stern (2019) sowie von Glymour (2015) vorgebracht.

shalb eine formale Erweiterung von Strukturgleichungsmodellen vor. Diese Erweiterung macht es möglich, Kausalbeziehungen auf mehreren Supervenienzebenen zu formulieren und erhöht damit deutlich die Ausdrucksmöglichkeiten innerhalb eines gegebenen Strukturgleichungsmodells.

In Kapitel 10 nutze ich schließlich den im vorhergehenden Kapitel erweiterten formalen Rahmen von Strukturgleichungsmodellen, um Yablo's (1992) Kriterium der Proportionalität zu explizieren. Ich zeige allerdings, dass dieses Kriterium nicht immer das ideale Level für eine kausale Erklärung auswählt. Ich expliziere daher ein neues Kriterium der Proportionalität (*MS-proportionality*), das zum Ziel hat, vor allem für solche potentielle Erklärungen das sinnvollste Level der Abstraktheit zu bestimmen, die als Kandidaten in einem Schluss auf die beste Erklärung fungieren.

Damit ist dann auch das dritte Hauptziel der Dissertation erreicht: Eine formal präzise Explikation eines zweckmäßigen und intuitiv adäquaten Kriteriums der Proportionalität einer Erklärung.

In der *Conclusion* fasse ich schließlich die wesentlichen Ergebnisse meiner Dissertation zusammen, weise auf einige Einschränkungen meiner Ergebnisse hin und zeige auf, wie zukünftige Forschung an meine Ergebnisse anknüpfen könnte, vor allem in Hinblick auf eine Formalisierung von Schlüssen auf die beste Erklärung.