KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Actual Causation: Definitions and Principles

Sander Beckers

Supervisors:
Prof. dr. ir. H. Blockeel
Prof. dr. J. Vennekens

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

October 2016

Actual Causation: Definitions and Principles

Sander BECKERS

Examination committee:
Prof. dr. ir. H. Van Brussel, chair
Prof. dr. ir. H. Blockeel, supervisor
Prof. dr. J. Vennekens, supervisor
Prof. dr. M. Denecker
Prof. dr. G. De Samblanx
Prof. dr. ir. T. Goedemé
Prof. dr. J. Heylen
Prof. dr. J. Halpern
(Cornell University)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

October 2016

Preface

How small a thought it takes to fill a whole life!

–Ludwig Wittgenstein, *Culture & Value*

It hasn't been a whole life, but sometimes it did feel that way. One day I woke up and asked myself the question, "What does it mean to say that some event caused another?", and here I am, more than four years later, with this dissertation in front of me. Old people tend to warn young people "that time flies by", and young people tend to ignore old people when they do. Yet that description perfectly captures the passage of time as it occurred during my PhD. When I started, I was in my twenties, I lived in Brussels renting a small studio, I would attend parties and/or bars at least once a week, and did not know anyone in the academic community outside of Leuven. Today, I am in my thirties, I live in the United States while owning an apartment in Brussels, I attend lectures and/or receptions at least once a week, and have made academic connections with people from all over the world. Still, the most fundamental change in my life during this period has been the writing of this text. As a young boy I already cherished vague dreams of some day joining the ranks of the "thinkers", i.e., those whose labour is visible only by reproducing their efforts in one's own mind. Whatever purpose this PhD dissertation may serve in aiding others with an interest in its topic, if nothing else it fulfills the promise that I made to that young boy.

In doing so, I also fulfilled my promise towards the kind people of the Flemish Agency for Innovation by Science and Technology (IWT). Without the scholarship that they awarded me, it would have been financially impossible to exclude myself from the labour market for four years whilst continuing a comfortable life. For this reason I would like to thank them first and foremost.

It is a common practice to thank several people who assisted the PhD student on his journey, but in my case it is far more appropriate to thank one person several times. Although for all official purposes Hendrik Blockeel is to be considered my principal

supervisor, anyone who has witnessed the progress of this work knows that it was Joost Vennekens who performed that role. These few lines written here are vastly insufficient to express my gratitude towards Joost for his support. To say that this work could not have been *finished* without him should not come as a surprise, as that probably holds true for any good supervisor. However in my case it is no overstatement to say that this work could not even have *started* without Joost.

I met Joost about halfway through my Master's degree in Mathematics, for which I was writing a thesis under the supervision of Marc Denecker. After having struggled for a long time with finding the right topic that was suitable with regards to both my competencies as a mathematician, and Marc's competencies in communicating the expected work to be performed, Marc contacted Joost in a final effort to help me complete my thesis. He guided me towards a topic to work on, and from that moment onwards he took on the role as my thesis supervisor. Hence I have him to thank for obtaining my Mathematics degree.

However my thanks extends even further. We got to know each other, and after a while he encouraged me to apply for a research position at Campus De Nayer, again under his supervision. It stands beyond a doubt that I lacked the proper qualifications for the position at hand, and I would not have stood a chance of obtaining it through any orthodox bureaucratic application procedure. Yet Joost had faith that I would be able to meet the challenge, and convinced his colleagues to hire me. Hence I have him to thank for obtaining my first research position.

After a year the project I had been working on came to an end, and I started thinking about research proposals for a PhD. In philosophical circles I noticed a growing interest in the issue of formally studying causality, and that sounded like an ideal field to get involved in. By delving into the literature, I came across a paper by Daan Fierens, who I was surprised to discover worked at this very department. So I send him an e-mail asking if he would consider helping me find a suitable research question. In his reply, he directed me towards a colleague of his, who he claimed was far more knowledgeable about the subject, as he had written several articles on it. Of course by now you can guess to which colleague he was referring. Unbeknownst to the both of us, I had been investigating a subject matter on which Joost was an expert. The very next day when I came into the office, I said to Joost: "We have to talk." A month later, we had written a research proposal on the topic of actual causation, and the rest is history. Hence I have him to thank for obtaining my second research position as well.

All of the former does not even mention the thanks I owe Joost for his assistance during the actual PhD itself. Some might not have enjoyed having Joost as a supervisor. More specifically, anyone who expects his supervisor to arrange weekly meetings, who expects him to work out a concrete plan on how to proceed, expects to be reminded of important deadlines, or expects him to be in his office most of the time, would probably not welcome Joost as a supervisor. Fortunately I expected none of those

things. Likewise, some supervisors might not have enjoyed me as a student. More specifically, anyone who expects his student to send a weekly overview of what he has done, who expects his work to be structured in concrete goals and tasks, who expects a PhD student in computer science not to spend most of his time engaging with the philosophy literature, or expects him to be in his office more than once a month, would probably not welcome me as a PhD student. Fortunately Joost expected none of those things. In that sense, we were a perfect match.

But I believe it is fair to say that we were also highly compatible when it comes to the substance of our work. Basically, Joost is an AI researcher with an interest in philosophy, and I am a philosopher with an interest in AI. Given the strong opinions that both of us hold on many issues, it is nothing short of remarkable that we always ended up finding a suitable compromise in the long run. I cannot overestimate the influence that he has had on directing me towards the views expressed in this work. He was quick to point out flaws in my reasoning, but never without suggesting alternative solutions. Further, his experience in writing technically solid and precisely articulated articles has proven invaluable for improving my own writing. Most importantly, working with Joost has always been a pleasant and stimulating experience. Hence I have him to thank not only for his help in completing my PhD, but also for making the process of doing so enjoyable during all this time.

Further, I would like to thank Hendrik for helping me settle in at the department, and for his willingness to take up the official role as my supervisor. I also owe many thanks to Marc Denecker and Joe Halpern for very stimulating discussions on actual causation. That gratitude extends to the other members of the jury as well, for having taken the time to critically read this work, and offering helpful comments.

I also want to thank my colleagues, both of the KRR research group, and at Campus De Nayer. Usually such thanks is owed due to the pleasant working environment they create, including the many coffee breaks, lunches, and office discussions. However it would be unfair to restrict my gratitude to those events, as they rarely took place. I probably deserve the title of least visible member of the computer science department, as I hardly ever made it to the office. Therefore I take full responsibility for the lack of social engagement with my co-workers, and I thank them for accepting my overwhelming absence without ever making me feel unwelcome.

Lastly, I want to thank my friends and family for taking my word on it that what sometimes seemed to them like a four-year long holiday was actually a valuable contribution to the development of human knowledge. Their relentless efforts in confronting me with the harsh realities of the world outside academia have not been in vain, as it motivated me to secure my current academic position, and has strengthened my intention to never again leave academia. On a more serious note, I thank them for their infinite support and understanding. Working on a PhD is undoubtedly a great privilege granted to me by society, but at the same time one should not underestimate

the accompanying mental burden that it can give rise to every now and then. I owe it to my friends and family that I was always able to rely on them in times of need.

Sander Beckers

Ithaca, October 2016

Nomenclature

actual causation Relation between two events/omissions (x, y) that occur in a story, which holds iff x caused y. (Definitions 1, 3, 32, and 33, Section 3.3, 3.4, and 3.7.)

anti-symmetry Holds for a binary relation $R(x, y)$ iff $R(x, y) \Rightarrow R(\neg x, \neg y)$. (Corollary 2.)

asymmetry Holds for a binary relation $R(x, y)$ iff $R(x, y) \Rightarrow \neg R(\neg x, y)$. (Principle 4, Section 7.5.1.)

branch A branch of a probability tree in CP-logic, used to formalise a story. (Section 2.3.1.)

contributor Relation between two events/omissions (x, y) that occur in a story, which holds iff x contributed in a very minimal sense to bringing about y. (Section 6.3 and 7.4.1.)

counterfactual dependence Relation between two events/omissions (x, y) that occur in a story, which holds iff intervening on x would have prevented y from occurring. (Definitions 2 and 31, Sections 3.4.1, 6.2, and 6.7.)

CP-law Formal representation of an independent causal mechanism in CP-logic. (Section 2.3.1.)

CP-theory Set of CP-laws which together make up a causal model that is an appropriate formalisation of a small part of the world. (Section 2.3.1.)

default The value a variable takes if nothing acts upon it. (Section 2.3.1.)

deviant Any value a variable takes if it is acted upon. (Section 2.3.1.)

early preemption Common story used in the literature to gauge intuitions of actual causation. It involves two causal mechanisms, one which is successful in causing an effect, and another which was preempted before the effect occurred. (Figure 2.3, Table 3.1, Section 5.2.1 and 7.3.1, Examples 12 and 13.)

endogenous Property of variables that holds if the variable is internal to the causal model of interest. Each endogenous variable has a unique equation in structural equations modelling. (Section 2.2.)

exogenous Property of variables that holds if the variable is external to the causal model of interest. An exogenous variable does not have an equation in structural equations modelling, rather its value is stipulated. (Section 2.2.)

intrinsic Property of a CP-law as it appears in a branch when considering an effect x, which holds if the law definitely played a role in causing x. (Section 3.2.1.)

irrelevant Property of a CP-law as it appears in a branch when considering an effect x, which holds if the law definitely did not play a role in causing x. (Section 3.2.1.)

late preemption Common story used in the literature to gauge intuitions of actual causation. It involves two causal mechanisms, one which is successful in causing an effect, and another which was preempted by the occurrence of the effect. (Examples 2 and 9, Table 3.1.)

literal Atomic formula of the form $V_i = v_i$, where V_i is a variable and v_i is one of the values it can take on.

necessary Property of a CP-law as it appears in a branch when considering an effect x, which holds if the law occurs in every reduction of the story, and has the same effect. (Definition 5.)

neuron diagram Graphical representation of a causal model. (Section 2.4.)

normality ranking An ordering of states of affairs based on how normal they are in the given context. (Section 4.2.)

preempted A contributor that failed to be a producer. (Section 6.4.2.)

probability tree Formal representation of the possible stories that a CP-theory allows. (Section 2.3.1.)

producer Relation between two events/omissions (x, y) that occur in a story, which holds iff x brings about y. (Sections 3.4.2 and 6.4.)

reduction Simplification of a story containing an effect x, so that the causes of x are still present. (Section 3.7, Lemma 1.)

simple Property of a story, which holds if there are no limitations on the choices made for laws that are not necessary. (Definition 7.)

story An example of a causal scenario in which two events/omissions x and y occur. Informally a story is given as a short description in natural language, formally a story can be described as a branch (CP-logic), a neuron diagram together with the states of its neurons, or the combination of a causal model M and the values of its exogenous variables $\vec{U}$ (structural equations modelling).

structural equation Formal representation of a causal mechanism in structural equations modelling. (Section 2.2.)

switch Common story used in the literature to gauge intuitions of actual causation. It involves two causal mechanisms, one which is successful in producing an effect, and another which was preempted before the effect occurred, but would otherwise have certainly been successful in producing the effect. (Table 3.1, Section 5.2.2 and 6.5.)

symmetric overdetermination Common story used in the literature to gauge intuitions of actual causation. It involves two causal mechanisms which are each sufficient for causing an effect, that both occur at the same time and thus symmetrically overdetermine the occurrence of the effect. (Section 5.3.1, Example .)

timing Formal representation of the order in which events occur in a story, that can be added to extend structural equation modelling. (Definition 22.)

transitivity Holds for a binary relation $R(x, y)$ iff $R(x, y) \wedge R(y, z) \Rightarrow R(x, z)$. (Chapter 7, Principle 5.)

Contents

Chapter 1

Introduction

This work is a study of the concept of *actual causation*. More specifically, the aim of this text is to develop formal definitions of actual causation that capture its informal meaning. We will do so from within two different perspectives. In Part I we use CP-logic to address actual causation from a practical perspective, meaning that the focus lies on quantifying the notion of actual causation, and adding extensions that allow taking into account the context-sensitivity of causation. In Part II we use structural equations modelling, and look at actual causation from a purely theoretical perspective. Concretely, we develop formal principles which a suitable definition of actual causation should satisfy, and present a definition that does so.

Actual causation, a.k.a. token causation, should be contrasted with *general causation*, a.k.a. type causation. Put briefly, actual causation is a relation between particular events, whereas general causation is a relation between types of events. We will have very little to say about the latter, but we can illustrate the difference between the two with a simple example.

In the statement "Smoking is a cause of lung-cancer", causation is to be interpreted as general causation: there are laws of nature such that many people who smoke will develop lung-cancer as a result. A general causal law is useful in predicting what will happen given that certain events occur. Therefore it is useful *before* we decide how to act. For example, we try to discourage people from smoking precisely because in general this causes lung-cancer.

On the other hand, in the statement "John's lung-cancer was caused by smoking", causation figures as actual causation: in the particular case of John, his lung-cancer was the result of him smoking (rather than, say, working in a coal mine). Actual causation is a relation that we can establish only *after* the facts have occurred, and only given that

we know all the relevant general causal laws. Note that the statement about smoking as an actual cause does not follow from the statement about smoking as a general cause: it is not because smoking is generally a cause of lung-cancer, that for every single person who both smokes and has lung-cancer, the latter was caused by the former.

The distinction between general and actual causation implies a first note of clarification. Our study is not an *empirical* inquiry regarding what causal laws can be discovered, but it is a *conceptual* investigation regarding the meaning of statements involving actual causation. Concretely, we assume all data that is relevant to a particular question of actual causation is given, and all the relevant general causal laws are known. In terms of our example from above, we assume that all details of John's physical condition are given, and we assume that all the relevant physiological mechanisms regarding lung-cancer are known. In other words, our study is an enquiry into the necessary and sufficient conditions for saying that some event C is an actual cause of some event E.

A second note of clarification is that we do not restrict actual causation to events, but also include *omissions*. An omission is here understood as the absence of an event, i.e., the non-occurrence of an event. For example, rather than stating that the event "Billy showed up in class today" did not occur, we can state the omission "Billy did not show up for class today".

A third note of clarification is that in this work we only consider so-called *counterfactual* approaches to actual causation. Other significant approaches (such as regularity based accounts, probability-raising accounts, or process-theories of causation) will not be discussed. The reason for this is that the former have proven far more successful in dealing with difficult examples, and are more readily expressed using the formal tools commonly found in the Artificial Intelligence literature. As a result, these types of approaches have received far more attention over the last few years and are in fact being applied in a wide range of domains going from legal reasoning, to fault diagnosis in software systems, to history, medicine, etc.

Although the concept of actual causation is rather technical, and our analysis will share that feature, it is one with which we are all very familiar: we use it to offer explanations.

In the overwhelming majority of cases where one asks the question: "Why did this event happen?", the answer given will refer to the actual causes of that event. To be sure, the concepts of explanation and actual causation – or causation, for short – are not entirely reducible to one and other, and the precise relation between them is an interesting subject matter in its own right. (For example, in some cases, the answer to the question posed might be a mathematical argument, or a piece of information regarding the relevant causal laws, or possibly another form of information that does not consist of actual causes.)

Nevertheless it will prove helpful to keep in mind the close connection between them, because the concept of "explanation" is far more familiar than that of "causation".

More specifically, throughout this work we will often present small, simple stories, and ask of two events – for which we shall use the letters C and E – that occur in this story, whether or not the one is a cause of the other. An important tool in answering that question is to consider whether or not the answer to the question "Why did E happen?" should contain a reference to C.

This brings us to the issue of methodology. On what grounds shall we decide if a suggested definition of causation is correct?[1] The previous paragraph suggests one important source of insight: common intuitions regarding everyday usage of the term. In some cases intuitions are universal and very strong, so that any definition which violates them ought to be immediately dismissed.

But of course intuitions can only get you so far. Very often they are not shared by everyone, or they are too vague, or simply inconsistent. In these situations, intuitions serve mainly as a source of inspiration, but do not have the last word. Rather, we aim to build up a definition of causation based on a small set of fundamental principles that causation satisfies.

Still, although in the end this work will exhibit a strong preference for one particular definition of causation based on such principles, we consider it safest to adopt a pluralist perspective, i.e., we will develop and defend several definitions of causation.

1.1 Structure of the text

Chapter 2 introduces the required formal tools that will be used in our investigations. Concretely, it introduces the formal languages of CP-logic and structural equations, and briefly discusses neuron diagrams. Further, it presents a method for translating between these three.

The remainder of the work is divided into two parts, the first consisting of Chapters 3 through 5, and the second consisting of Chapters 6 and 7.

Rather than focussing on one particular definition, in Part I we use the formal language of CP-logic to develop a general framework that allows for several definitions of causation, all of which offer a probabilistic degree of causation, and can be extended in order to incorporate context-sensitive factors regarding normality. Concretely, Chapter 3 develops a general parametrised definition of causation, and Chapter 4 shows how this definition can be extended to incorporate the influence of normality.

The focus in this part lies on searching definitions of causation that lend themselves to a wide variety of applications. The main result will be two definitions, Hall07 and BV12, which are both compromises between the concepts of *dependence* and *production*,

[1] For a more elaborate analysis of this issue, see (Paul & Hall, 2013)[Ch. 2].

defined here as well. Unfortunately both definitions have certain shortcomings, which we discuss in Chapter 5.

Part II focusses entirely on finding a single, binary, definition of actual causation that is founded on suitable underlying principles, and does so using the language of structural equations. (We point out that we do extend structural equations with a *timing*, in order to mimic the temporal information captured by the semantics of CP-logic.)

Chapter 6 discusses several paradigmatic examples that lead the way to four important principles underlying causation, and a definition of causation that satisfies them. Chapter 7 complements this discussion by analysing the importance of one controversial principle, namely the transitivity of causation. This analysis ends up offering further support for the definition developed in Chapter 6.

Although the resulting definition corrects the shortcomings presented in Chapter 5, all three definitions here presented agree on a majority of examples. Further, similar to the two definitions from Part I, the definition from Part II is also a compromise between *dependence* and *production*, be it that the concept of production used here is more general than its counterpart above.

Chapter 2

Formal Background

2.1 Introduction

Ever since Lewis (1973) first analyzed the problem of actual causation in terms of counterfactual dependence, philosophers and researchers from the AI community alike have been trying to improve on his attempt. The literature now contains many definitions of actual causation, each with its own strengths and weaknesses. The fact that the formulations of these different definitions diverge widely, proves a major obstacle for evaluating and comparing them. A second problem is posed by the standard practice of assuming that questions of causation can be separated from their field of application, because this conflicts with recent research that suggests actual causation is a strongly context-dependent concept. In this first part we address both issues, by presenting a formal framework for expressing definitions and extensions of actual causation.

Following Pearl (2000), structural equations have become a popular formal language for defining actual causation (Halpern, 2015a; Halpern & Pearl, 2005a; Hitchcock, 2007, 2009; Weslake, 2015; Woodward, 2003). A notable exception is the work of Ned Hall, who has extensively critiziced the privileged role of structural equations for causal modelling, as well as the definitions that have been expressed with it. He has proposed several definitions himself (Hall, 2004, 2007; Hall & Paul, 2003), the latest of which is a sophisticated attempt to overcome the flaws he observes in those that rely too heavily on structural equations. Building on a definition by Vennekens (2011), we have developed a definition of our own (Beckers & Vennekens, 2012) within the language of CP-logic (Causal Probabilistic logic). CP-logic (Vennekens, Denecker, & Bruynooghe, 2009) is a probabilistic logic programming language, based on Sato's

distribution semantics (Sato, 1995).

The relation between these different approaches is currently not well understood. Indeed, they are all expressed using different formalisms (e.g., neuron diagrams, structural equations, CP-logic, or just natural language). Therefore, comparisons between them are limited to verifying on which examples they (dis)agree. Our first goal in this part is to work towards a remedy for this situation. We will present a general, parametrized, and probabilistic definition of actual causation.

We will develop this definition in the context of CP-logic, because this language offers all the features that are required to define the necessary concepts in a straightforward and natural way. In particular, we make use of the fact that CP-logic has a modular rule-based structure and a semantics that is explicitly temporal and makes a distinction between the default and deviant values of variables. In the context of structural models, which lack these features, our general framework would be significantly more cumbersome to define.

Exploiting the fact that neuron diagrams and structural equations – at least those typically used in the actual causation literature – can be reduced to CP-logic, we will then show that our definition and three definitions by Ned Hall can be seen as particular instantiations of this parametrized definition. Our analysis thus allows for a formal and fundamental comparison between these approaches, which is a first step towards an improved account. Also, it generalises the definitions by Hall from a deterministic to a non-deterministic setting. Still, all of these definitions share one feature: they do not take into account the context in which the question of actual causation is posed.

Halpern and Hitchcock (2015) draw on empirical research regarding people's causal judgments, to suggest a graded and context-sensitive notion of actual causation. Concretely, they propose an *extension* to actual causation that takes into account the normality of different settings. Although we sympathise with many of their observations, their restriction to a merely partial normality ordering runs into trouble for more complex examples. Therefore we aim to improve on their approach, by defining an extension to our general definition of causation, completing our general framework for the study of actual causation.

In the next section we introduce structural equations, and present the definition of actual causation the success of which has made them such a popular tool, namely that of Halpern and Pearl (2005a). This is followed by an introduction to the CP-logic language in Section 2.3. Since Hall presents his examples using neuron diagrams, we also briefly discuss these in Section 2.4, followed by a method for translating causal models between these three different formal representations.

With those preliminaries out of the way, Chapter 3 presents a general definition of actual causation using CP-logic. This definition is then instantiated into four concrete definitions. We offer a succinct representation of all these definitions, and an illustration

of how they compare to each other. The second goal of this first part is addressed in Chapter 4, where we focus on the extension to actual causation by Halpern and Hitchcock. First we translate their work into the CP-logic language, after which we offer several improvements, and a generalisation to allow for non-deterministic examples. In Chapter 5 we present some problems for the definitions developed so far, which leads the way to an improved approach to actual causation in the second part.

2.2 Structural Equations Modelling

We briefly introduce a simple version of structural equations modelling, which is the most popular formal language used to represent causal models. In general, structural equations allow functional dependencies between continuous variables, or discrete variables with possibly an infinite domain. However, the actual causation literature typically considers only examples made up of discrete variables with a finite domain, and propositional formulas. Further, in the majority of cases the variables are Boolean. This is why we restrict attention to those kinds of models. For a detailed introduction, see (Pearl, 2000).

A structural model consists of a set of *endogenous* variables $\vec{V}$, a set of *exogenous* variables $\vec{U}$, and a causal model M. Although we only consider models with Boolean variables, we should point out that the results we will present can easily be generalized to allow for multi-valued variables as well. We explain this below.

The model M is a set of *structural equations* so that there is exactly one equation for each variable $V_i \in \vec{V}$. An equation takes the form $V_i := \phi$, where ϕ is a propositional formula over $\vec{V} \cup \vec{U}$. For any variable V_i, we denote by ϕ_{V_i} the formula in the equation for V_i in M. We follow the customary practice of leaving the equations for variables that depend directly on the exogenous variables implicit, and simply state the value they take in each particular story.

For an assignment $(\vec{v},\vec{u})$ of values to the variables in $\vec{V} \cup \vec{U}$, we denote by $\phi^{(\vec{v},\vec{u})}$ the truth value obtained by filling in the truth values $(\vec{v},\vec{u})$ in the formula ϕ. An assignment $(\vec{v},\vec{u})$ *respects* M, if for each endogenous variable V_i, its value $v_i = \phi_{V_i}^{(\vec{v},\vec{u})}$. As usual, we only consider models M in which the equations are acyclic, which implies that for each assignment $\vec{u}$ to $\vec{U}$, there is exactly one assignment $(\vec{v},\vec{u})$ that respects M. Therefore, we refer to $\vec{U} = \vec{u}$ as a *context*. For every value $\vec{u}$ of $\vec{U}$, we call the pair $(M,\vec{u})$ a *causal setting*. We write $(M,\vec{u}) \models \phi$ if $\phi^{(\vec{v},\vec{u})} = \textbf{true}$ for the unique assignment $(\vec{v},\vec{u})$ that respects M.

A *literal* L is a formula of the form $V_i = v_i$ or $U_i = u_i$. Our restriction to Boolean variables is made concrete here: the only values v_i and u_i we consider are **true** and

false. Hence our definitions and results can be generalised by simply lifting this restriction. The same comment applies to CP-logic, introduced in the next section.

We will use the atom V_i as a shorthand for $V_i = \textbf{true}$, and the negated atom $\neg V_i$ as a shorthand for $V_i = \textbf{false}$. Regardless of whether $L_i \equiv V_i$ or $L_i \equiv \neg V_i$, we write ϕ_{L_i} to mean ϕ_{V_i}. Hence in the case where $L_i \equiv \neg V_i$, $\neg\phi_{L_i}$ will be a propositional formula that makes L_i true in any assignment that respects M. Further, we denote by $L_{(M,\vec{u})}$ the set of all literals L_i such that $(M,\vec{u}) \models L_i$. Further, we denote by $L_{(M,\vec{u})}$ the set of all literals L_i such that $(M,\vec{u}) \models L_i$.

A causal model M is a tool to represent *counterfactual* relations between variables, in the sense that changing the values of the variables on the right-side of an equation can change the value of the variable on the left-side, but not vice versa. This makes them suitable devices to model *interventions* on an actual setting, meaning changes to the value of a variable V_i that affect only the values of variables that depend on V_i, but not those on whom V_i itself depends.

Syntactically, we make use of the $do()$-operator introduced by Pearl (2000) to represent such an intervention. For a model M and an endogenous variable V_i, we denote by $M_{do(V_i)}$ and $M_{do(\neg V_i)}$ the models that are identical to M except that the equations for V_i are $V_i := \textbf{true}$ and $V_i := \textbf{false}$, respectively. Hence for a causal setting $(M,\vec{u})$ such that $(M,\vec{u}) \models C$, the causal setting $(M_{do(\neg C)},\vec{u})$ corresponds to the counterfactual setting resulting from the intervention on $(M,\vec{u})$ that prevents C. We generalise the $do()$-operator to a set of literals in an obvious way, namely by replacing the equation for each literal in the set in the manner explained above.

2.2.1 HP-definition of Actual Causation

A detailed discussion of the HP-definition from (Halpern & Pearl, 2005a) is delayed until Part II. We already present it here because we will make reference to it every now and then throughout this text. We here present a somewhat simplified version. Concretely, we only consider single literals as causes or effect, as opposed to Boolean combinations of literals.

Definition 1 (HP definition of actual causation)**.** *Given $(M,\vec{u}) \models C \wedge E$, we define C to be an* actual cause *of E w.r.t. $(M,\vec{u})$ if there exists a partition $(\vec{Z},\vec{W})$ of $L_{(M,\vec{u})}$ with $C \in \vec{Z}$ such that both of the following conditions hold:*

1. *$(M_{do(\neg C,\neg\vec{W})},\vec{u}) \models \neg E$. In words, changing the actually satisfied literals $C, W_1, \ldots, W_n$ to their negation changes E from* **true** *to* **false**.
2. *$(M_{do(C,\neg\vec{W}',\vec{Z}')},\vec{u}) \models E$ for all subsets $\vec{W}'$ of $\vec{W}$ and all subsets $\vec{Z}'$ of $\vec{Z}$. In words, changing the values of any subset of literals in $\vec{W}$ should have no effect on E, as*

long as C is held fixed, even if all the literals in an arbitrary subset of $\vec{Z}$ are set to their original values in the context $\vec{u}$.

Here, the interventions $do(\neg\vec{W})$ represent changes to the actual setting beyond an intervention on the putative cause C, and are referred to as *structural contingencies*. The idea is, roughly, that $do(\neg C)$ is not sufficient to guarantee the truth of $\neg E$ by itself, and other interventions may be required, but C is sufficient to make sure E holds in the current context.

2.3 CP-logic

CP-logic (short for Causal Probabilistic logic) was originally introduced by Vennekens et al. (2009) as an expressive causal modelling language. Although there is much to say in favour of expressivity in general, for the current investigations it will be helpful to have a little less of it. Specifically, CP-logic is designed to be a useful tool for modelling *general* causal relations, rather than *actual* causal relations. Therefore we shall only use a fragment of CP-logic throughout this work.

2.3.1 Formal Semantics of CP-logic

A CP-law

The basic syntactical unit of CP-logic is a *CP-law*, which takes the general form:

$$(A_1 : \alpha_1) \vee \ldots \vee (A_n : \alpha_n) \leftarrow \phi$$

Here each A_i is a Boolean variable, or an *atom*, each α_i belongs to $[0,1]$, and $\sum \alpha_i \leq 1$. ϕ is a formula, the form of which we describe below. We refer to the part to the left of the arrow as the *head* of the CP-law, and to the right as its *body*.

A CP-law represents an *independent, non-deterministic, causal mechanism*: the body represents the cause for the mechanism to be "activated", or "applied", and the head represents the probability distribution over all its possible effects. Thus, each A_i represents an *event* that can be the result of the mechanism. The reason we do not demand $\sum \alpha_i = 1$, is that we want to allow for a mechanism having no effect at all. Another way to interpret this is by considering the head to contain an *empty disjunct*, representing the lack of an effect. The probability for this disjunct is then given by $(1 - \sum \alpha_i)$.

In CP-logic in general, ϕ may consist of any first-order logic formula. That we are considering only a fragment of CP-logic, is made concrete by the fact that we only

consider ϕ as a conjunction of ground literals, i.e., a conjunction of atoms and negated atoms. The motivation behind this restriction goes in two steps.

First, actual causation is taken to be a relation between an *event* or an *omission*, and another *event* or *omission*, and those are represented by literals. More specifically, we represent an event by an atom, and an omission, i.e., the non-occurrence of an event, by a negated atom. Complex events are then represented by propositional combinations thereof. This offers a first restriction: we only consider propositional formulas ϕ.

Second, we noted that a CP-law represents an *independent* mechanism. To ensure this independence, we do not consider "disjunctive mechanisms" – meaning a mechanism that can be triggered by either one of two causes – as primitives of the language. Instead, we model such behaviour by two separate mechanisms, which taken together can sometimes behave as a disjunctive mechanism, a topic we will discuss in detail later. Therefore as a second restriction we exclude all disjunctions from ϕ.

Concretely, a CP-law takes the general form:

$$(A_1 : \alpha_1) \vee \ldots \vee (A_n : \alpha_n) \leftarrow B_1 \wedge \ldots \wedge B_m$$

Here each B_j is either an atom or a negated atom, meaning it represents either an event or an omission. In practice, every CP-law in all examples discussed in this work will contain just a single atom A in its head. If the corresponding probability $\alpha < 1$, then such a law can be read as stating: if the mechanism is activated, it will either produce A or will have no effect at all.

It is common in causal modelling to divide all variables into *endogenous* and *exogenous* ones. Exogenous variables are taken to represent the background conditions, or the *context*, of the local causal model under consideration, meaning that for any actual story their values are simply given. The endogenous variables on the other hand represent the variables whose causal relations we are interested in. Some of these endogenous variables are determined directly by the context, i.e., by the exogenous variables. The equivalent of such a variable in CP-logic is a variable appearing in the head of a law whose body is $\top$. (In such cases we shall simply omit the body.) Concretely, to say that A is determined directly by the exogenous variables, is translated into CP-logic as the law $(A : \alpha) \leftarrow$.

Assume we are considering a CP-law in isolation. If the body is satisfied, then the mechanism is triggered, and it produces one of the effects A_i (or has no effect). We call this process an *atomic story*, and consider each B_j to be an *atomic actual cause* of the effect A_i. We take this atomic relation as primitive, not in need of any further analysis. This means our aim is not to give a reductive account of actual causation, where actual causation is reduced to some other more primitive relation. Rather, the aim is to construct a definition of actual causation in complex stories, where multiple mechanisms interfere with one and other, in terms of actual causes in these atomic

stories. In this manner we sidestep the controversial metaphysical debate on the ultimate nature of causation.

We should point out that it is quite common in the counterfactual tradition for an account of causation to be non-reductive in this sense.[1] Consider for example the influential account by Halpern and Pearl (2005a)[p. 5]:

> It may seem strange that we are trying to understand causality using causal models, which clearly already encode causal relationships. Our reasoning is not circular. Our aim is not to reduce causation to noncausal concepts, but to interpret questions about causes of specific events in fully specified scenarios in terms of generic causal knowledge such as what we obtain from the equations of physics.

A CP-theory

A *CP-theory* T is a finite set of CP-laws, and represents a local causal model of the relevant part of the world under consideration. The stories of interest to us are the complex stories described by such theories, and actual causation is a relation between events/omissions in such stories. Informally, an *actual story* is a sequence of events that occur (or fail to occur) according to the mechanisms captured by a theory T. A theory is a general description of reality, in that it allows for a range of different possible stories. There are two degrees of freedom offered by T. First, a non-deterministic law allows for different effects, each occurring with the corresponding probability. Second, the order in which mechanisms are activated can vary.

We illustrate with a variant of an example from (Hall, 2004):

Example 1. *Suzy and Billy can both decide to throw a rock at a bottle. When Suzy does so, her rock shatters the bottle with probability* 0.9*. Billy's aim is slightly worse and he only hits with probability* 0.8*.*

This small causal domain can be expressed by the following CP-theory T:

$$(Suzy : *) \leftarrow . \qquad (2.1) \qquad\qquad (BS : 0.9) \leftarrow Suzy. \qquad (2.3)$$

$$(Billy : *) \leftarrow . \qquad (2.2) \qquad\qquad (BS : 0.8) \leftarrow Billy. \qquad (2.4)$$

Suzy (resp. *Billy*) represents Suzy (resp. Billy) throwing a rock, and *BS* represents the bottle shattering. Laws (2.1) and (2.2) state that *Suzy* and *Billy* are determined directly

[1] See (Paul & Hall, 2013)[Ch. 3] for a detailed discussion of this topic.

by the context. We use the notation $*$ to indicate that the precise value of the probability is not of interest, except that it is strictly smaller than 1.

Considered by itself, law (2.3) states that, if Suzy throws her rock, it will shatter the bottle with probability 0.9. Due to the presence of (2.4), this interpretation cannot be applied here. Instead, the meaning of (2.3) here involves another conditional: if Suzy throws her rock, *and the bottle has not yet shattered when it arrives at the bottle's location*, it will shatter the bottle with probability 0.9. (Note that both interpretations are equivalent in the absence of law (2.4).) The same obviously holds for law (2.4), be it that the probability is 0.8. Thus the fundamental temporal property that causes come before their effects is built directly into the semantics of a CP-law.

To formalize these ideas, the semantics of CP-logic uses *probability trees* (Shafer, 1996). The basic idea is that such a tree $\mathscr{T}$ represents possible evolutions of the domain from an initial state into a final state. For this example, one such tree is shown in Figure 2.1. Here, each node x is mapped to an interpretation of the variables $\mathscr{I}(x)$, which represents the corresponding state of the domain. (The white nodes indicate interpretations where the effect, *BS*, is **true**.) In the initial state of the domain (the root node), all variables are assigned their *default* value **false**: no event has occurred yet. In this example, the bottle is initially unbroken and the rocks are still in Billy and Suzy's hands.

The children of a node x are the result of the activation of a law r, whose body is satisfied in $\mathscr{I}(x)$: each edge (x,y) corresponds to a specific disjunct that was chosen from the head of r. The fact that a node has different children corresponds to the first degree of freedom, namely the non-determinism. Each node is labelled with the law r that is activated, and each edge is labelled with the chosen disjunct and its corresponding probability. As in the head of a CP-law, we leave implicit the empty disjunct representing the lack of an effect, since its probability can be derived from the others: the sum of the labels of all outgoing edges from a non-leaf node must always be 1.

Of course there may be several candidate laws that could be activated in x. For each candidate law, there will be a separate probability tree in which it is activated in x. This corresponds to the second degree of freedom, namely the order of the mechanisms being active.

When no more laws can be activated, a branch ends. The resulting leaf node represents the end state of the domain. In this manner each branch of a probability tree represents a possible *story* that might take place, according to the laws of the CP-theory. For example, our theory allows for the following story, which is a classic case of *Late Preemption*:

Example 2 (Late Preemption)**.** *Suzy and Billy both throw a rock at a bottle. Suzy's rock gets there first, shattering the bottle. However Billy's throw was also accurate,*

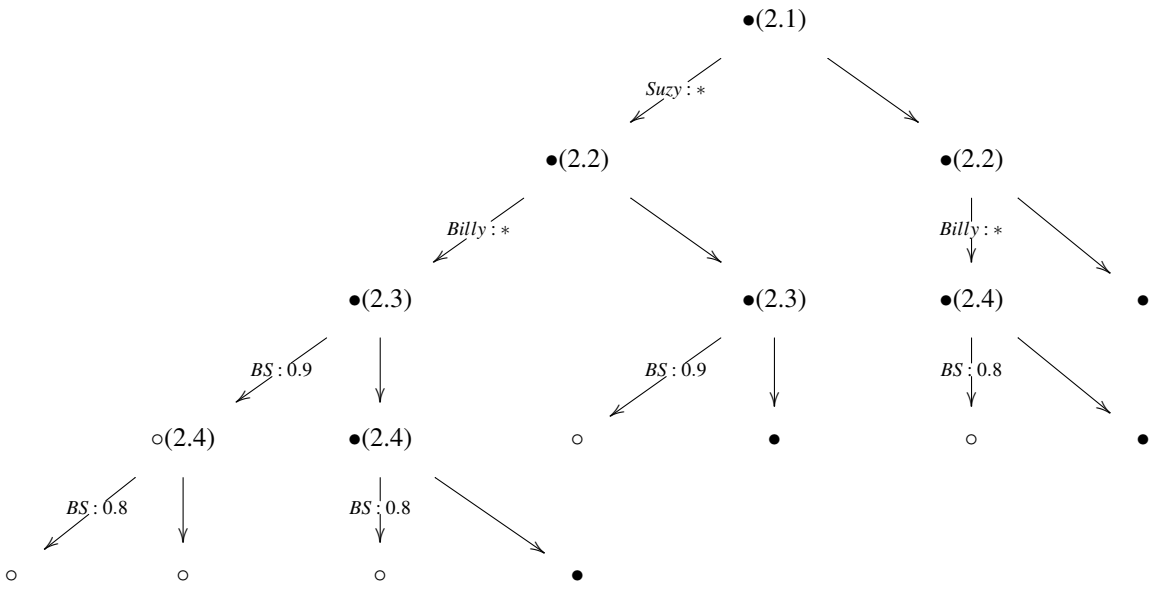


Figure 2.1: Probability tree for Suzy and Billy.

and would have shattered the bottle had it not been preempted by Suzy's throw.

The formal counterpart of this story is the leftmost branch b of our tree. We denote the n-th node of a branch b by n_b. As always, in the root node all variables are false: $\mathscr{I}(1_b) = \{\neg Suzy, \neg Billy, \neg BS\}$. Law (2.1) is activated first, so the state of the domain in 2_b, the left child-node of the root node, is obtained by setting *Suzy* to **true**, its *deviant* value. This is followed by law (2.2), so in $\mathscr{I}(3_b)$ we have that both *Suzy* and *Billy* are **true**. Then law (2.3) is activated and the bottle is broken, meaning *BS* is set to **true**. Finally, the last edge represents the fact that Billy's throw was also accurate, even though there was no bottle left to break. Hence $\mathscr{I}(4_b) = \mathscr{I}(5_b) = \{Suzy, Billy, BS\}$.

Each probability tree $\mathscr{T}$ defines an obvious probability distribution $\mathbf{P}_{\mathscr{T}}$ over its leaves, namely, the probability $\mathbf{P}_{\mathscr{T}}(l)$ of a leaf l is the product of the probabilities of all edges that lead to l. Given that each leaf is mapped to an interpretation of the domain, the probability distribution over leaves of the tree induces an obvious probability distribution over interpretations (the probability of I is $\sum_{\mathscr{I}(l)=I} \mathbf{P}_{\mathscr{T}}(l)$) and over Boolean formulas (the probability of ϕ is $\sum_{\mathscr{I}(l)\models\phi} \mathbf{P}_{\mathscr{T}}(l)$).

As we noted, there may be several laws that can be activated in a node x, giving rise to several probability trees. For example, the tree in Figure 2.1 only contains stories

where Suzy throws or does not throw first. An important property however is that all trees defined by the same theory result in the same interpretations in the leaf nodes, and thus also the same probability distribution. This is because if a law can be activated in some node x, then it can be activated in all subsequent nodes as well. This property is obvious for CP-laws whose precondition does not contain negation, because subsequent interpretations only increase in truth, i.e., the only thing that happens is that more and more variables deviate from their default value. The semantics of CP-logic takes special measures to ensure that this same property also holds for CP-laws containing negation. This is done by ensuring that a CP-law whose precondition depends on some variable V still being in its default state can only be activated once there no longer exists any way in which V could still be caused to deviate. In other words, it is not enough that V is false in the current state x, but it must actually be the case that V has already become impossible. This is formally defined by means of a fixpoint construction that overestimates everything that is still possible in x. We refer the reader to Vennekens et al. (2009) for the details of this construction. The bottom-line, however, is that it produces an interpretation $\mathscr{U}(x)$ that assigns **true** to all variables for which it is still possible that they could become true in some descendant of x. A law may then only be activated in x, if its body holds in both $\mathscr{I}(x)$ and $\mathscr{U}(x)$, since this means that it is not only true now, but will remain true from now on.

This ensures that, as far as the leaf nodes are concerned, the order in which CP-laws are activated does not matter. All probability trees $\mathscr{T}$ that can be constructed define precisely the same probability distribution $\mathbf{P}_{\mathscr{T}}$. For a CP-theory T, we denote this unique distribution by $\mathbf{P}_T$.

When it comes to actual causation, however, the order does matter. As mentioned, the different orders in different trees correspond to the different orders in which events can happen in a causal story. Consider again *Late Preemption*. Here it is intuitively evident that Suzy's throw caused the bottle to shatter, and Billy's throw did not.

Now imagine the example were slightly different, so that Billy's rock actually hit the bottle first. In that case, our judgment would be reversed: Billy's throw caused the bottle to shatter, and Suzy's throw did not. The formal counterpart of the story would be the leftmost branch of the probability tree depicted in Figure 2.2. (Note that if we were to reverse the order of the first two edges, then this would represent the story where Suzy throws before Billy, but Billy throws harder and thus his rock still reaches the bottle first.) Given that the interpretations in the leaf nodes of the branches for both stories – the leftmost branch in both Figure 2.1 and 2.2 – are identical, taking into account the order of the CP-laws is necessary to distinguish the causes in both stories. (In the next chapter we illustrate this by means of the definition from Section 3.3.)

Probability trees do not allow for simultaneous events: at every moment only one event can happen. Although this is a realistic assumption when considering a continuous time-scale, the granularity of the variables of interest in typical examples is usually

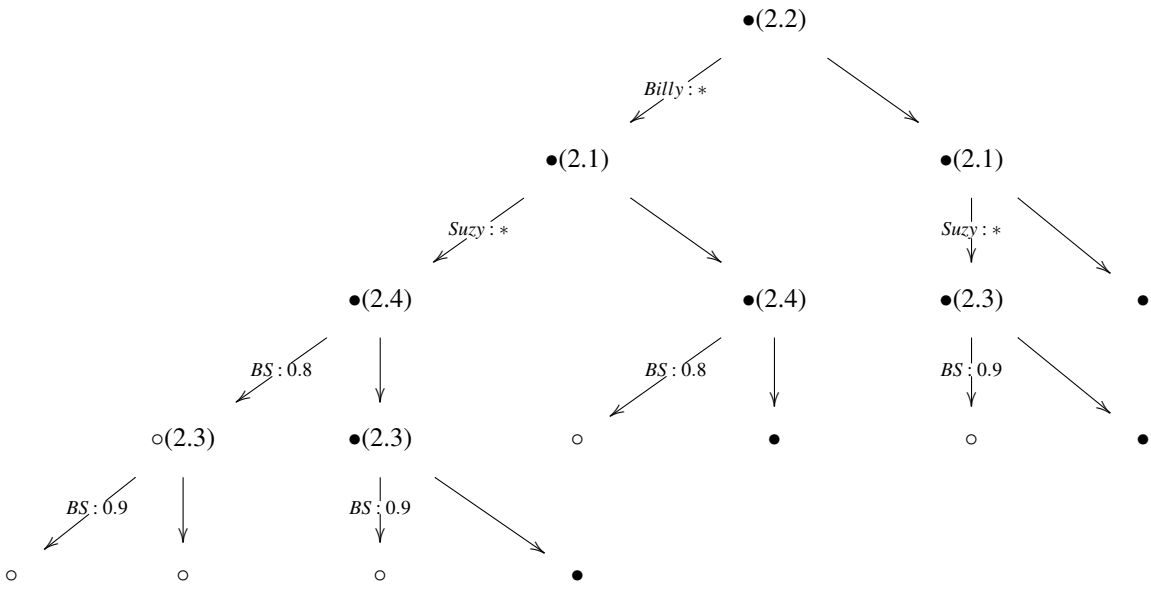


Figure 2.2: Another probability tree for Suzy and Billy.

closer to a coarser, discrete division of time. More concretely, we can easily imagine a variant of the story so that we can't distinguish which rock hits the bottle first. In that case it makes perfect sense to assume that for all intents and purposes both rocks hit the bottle simultaneously, making it a case of *Symmetric Overdetermination*. To accommodate this possibility we take such a story to correspond to two branches rather than one. These branches consist of the same nodes, ordered differently (eg., the leftmost branches of the trees in Figures 2.1 and 2.2 together represent the story in which Suzy's and Billy's rocks hit the bottle at the same time). We shall say that C causes E if this holds in either one of the branches considered by themselves.[2]

2.3.2 Counterfactual Probabilities

In the context of structural equations, Pearl (2000) studies counterfactuals and shows how they can be evaluated by means of a syntactic transformation. In their study

[2]We choose to define causation for these cases in this manner because it agrees with the majority verdict in the literature that both overdetermining events are causes. However one could also define it so that there has to be causation in both branches. In this case, neither event would be considered a cause. If there are more than two simultaneous events, we generalize this reasoning in the straightforward way.

of actual causation and explanations, Halpern and Pearl (2005b, p. 27) also define counterfactual probabilities (i.e., the probability that some event would have had in a counterfactual situation). Vennekens, Denecker, and Bruynooghe (2010) present an equivalent method for evaluating counterfactual probabilities in CP-logic, also making use of syntactic transformations.

Assume we have a branch b of a probability tree of some theory T. To make T deterministic *in accordance with the choices made in b*, we transform T into T^b by replacing the heads of the laws that were applied in b with the disjuncts that were chosen from those heads in b. For example, if we take as branch b the story from Example 2, then T^b would be:

$$Suzy \leftarrow . \qquad BS \leftarrow Suzy.$$

$$Billy \leftarrow . \qquad BS \leftarrow Billy.$$

We will use Pearl's $do()$-operator to indicate an intervention (Pearl, 2000). The intervention on a theory T that ensures variable C remains false, denoted by $do(\neg C)$, removes C from the head of any law in which it occurs, yielding $T|do(\neg C)$. For example, to prevent Suzy from throwing, the resulting theory $T|do(\neg Suzy)$ is given by:

$$\leftarrow . \qquad (BS : 0.9) \leftarrow Suzy.$$

$$(Billy : *) \leftarrow . \qquad (BS : 0.8) \leftarrow Billy.$$

Laws with an empty head are ineffective, and can thus simply be omitted. The analogous operation $do(C)$ on a theory T corresponds to adding the deterministic law $C \leftarrow$.

With this in hand, we can now evaluate a Pearl-style counterfactual probability "given that b in fact occurred, the probability that $\neg E$ would have occurred if $\neg C$ had been the case" as $\mathbf{P}_{T^b}(\neg E|do(\neg C))$.

2.4 Neuron Diagrams, Structural Equations, and CP-logic

In a standard neuron diagram, a neuron can be in one of two states, the default "off" state and the deviant "on" state in which the neuron "fires". An edge represents the influence of the state of the neuron on the left on the state of the neuron on the right of the edge. If the edge ends in an arrow, eg., the edge from C to D below, then the

influence is positive: C firing contributes to D firing as well. Given that there are no other incoming edges in D, C firing is sufficient and necessary for D's firing. The presence of an edge ending in a bullet, as is the case between C and B, represents a negative influence: the firing of C prevents B from firing, regardless of the state that A is in. When there are multiple incoming edges with arrows tips, as with E, then each of the neurons firing is sufficient for E to fire, as long as there is no preventive edge with a bullet.

Concretely in Figure 2.3, E fires iff at least one of B or D fires, D fires iff C fires, and B fires iff A fires and C doesn't fire. Neurons that are "on" are represented by full circles and neurons that are "off" are shown as empty circles.

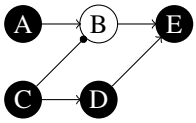


Figure 2.3

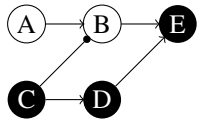


Figure 2.4

The diagrams in Figures 2.3 and 2.4 represent the same causal structure, but different stories: in both cases there are two causal chains leading to E, one starting with C and another starting with A. But in Figure 2.3 the chain through B is preempted by C, whereas in Figure 2.4 there is nothing for C to preempt, as A doesn't even fire. Therefore the first is an example of what is generally known as *Early Preemption*, whereas the second is not.

Although Hall (2007) presents his arguments using neuron diagrams, he formulates his definition in terms of structural equations that correspond to such diagrams in the following way: for each endogenous variable there is one equation, which contains a propositional formula on the right-hand side concisely expressing the dependencies of the diagram. The left-most neurons, which have no incoming edges, are determined directly by the exogenous variables $\mathbf{U}$ that represent the background conditions. For example, the structural equations for the diagrams in Figures 2.3 and 2.4 are:

$$A := U_1. \qquad B := A \wedge \neg C. \qquad E := B \vee D.$$

$$C := U_2. \qquad D := C.$$

We can translate such a structural model into an equivalent CP-logic theory by simply replacing the ":=" symbol by "$\leftarrow$", and replacing the exogenous variables by non-determinism in the laws (Vennekens et al., 2009). (Although we have restricted ourselves to Boolean variables, we point out that – just like structural equations –

CP-logic can be generalised to allow for multi-valued variables as well, as long as we stipulate a default value for each variable.)

CP-logic allows representations of causal relations that are more refined than those of structural models in two ways. First, cyclic causal relations can be represented in a more correct way than considered in, e.g., Pearl (2000) – for a discussion of this, see Vennekens et al. (2010)[Section 5].

Second, CP-logic is better suited for representing independent causal mechanisms: rather than having a single CP-law that combines all direct causes of a variable, CP-logic splits up independent mechanisms into separate CP-laws (Vennekens et al., 2010). Halpern and Pearl (2005a)[p. 2] also endorse such a modular approach for structural equations: "Each equation represents a distinct mechanism (or law) in the world, one that may be modified (by external actions) without altering the others." However, due to the static semantics of structural equations, all mechanisms influencing an endogenous variable have to be combined into one equation. As a result, a single structural equation sometimes represents more than one distinct mechanism.

For example, consider the theory from Section 2.3.1. There, law (2.3) represents the distinct mechanism that connects Suzy's throw to the bottle shattering $((BS : 0.9) \leftarrow Suzy)$, and likewise for law (2.4) and Billy's throw $((BS : 0.8) \leftarrow Billy)$. Using structural equations, one is forced to represent both mechanisms in a single law, using a disjunction to separate them: $BS := Suzy \vee Billy$. In this manner one loses the ability to distinguish between *Suzy* and *Billy* in the *Late Preemption* example we discussed earlier.[3]

Therefore we translate a single structural equation into a set of independent causal mechanisms. Likewise, we translate the influence of B and D on E in Figure 2.3 into two separate laws. Concretely, the translation of the causal model from Figures 2.3 and 2.4 into CP-logic is given by the following CP-theory – where p and q represent some probabilities:

$$(A : p) \leftarrow . \qquad B \leftarrow A \wedge \neg C. \qquad E \leftarrow B.$$

$$(C : q) \leftarrow . \qquad D \leftarrow C. \qquad E \leftarrow D.$$

In general, a CP-theory that is a translation of a neuron diagram contains two types of laws: those for the endogenous variables that are directly determined by the exogenous variables (i.e., the neurons with no incoming edges), and those for the downstream endogenous variables. By removing the exogenous variables and replacing it by non-determinism, the first type of law takes on the form $(A_i : p) \leftarrow$ for some atom A_i

[3] In Part II we discuss how Halpern and Pearl (2005a) try to work around this problem in order to deal with *Late Preemption* cases.

and a probability p. The second type of law is deterministic, and does not have an empty body: $A_j \leftarrow B_1 \wedge \ldots \wedge B_m$. Further, neuron diagrams cannot express *causation by omission* directly, i.e., they cannot express that the failure of a neuron to fire is sufficient for another neuron to fire. This means at least one of the literals B_i will be an atom (as opposed to a negated atom).

The state of the neuron diagram (i.e., which variables fire and which do not) corresponds to an assignment $\mathbf{v}$ of values to the Boolean variables $\mathbf{V}$ of the corresponding structural equations model M, which in turn corresponds to an interpretation for a vocabulary of the corresponding CP-logic theory. With such a state, we therefore associate a set of branches of probability trees of the CP-logic theory whose leaf nodes contain this interpretation. To construct this set, we first need to explain the temporal interpretation of a neuron diagram.

The direction from left to right in a neuron diagram should be interpreted temporally: neurons that are aligned vertically all fire or fail to fire at the same time point. Concretely, the set of all leftmost neurons represents the initial time point 0, and then each set of vertically aligned neurons to the right of the previous one represents the next time point. Hence for each neuron there is a single time point at which it can fire.

For example, in the diagram from Figure 2.3, both A and C fire at time point 0. Then, at time point 1, D fires. The fact that B does not fire at this time implies that it will never fire. Finally, E fires at the next time point. For ease of exposition, we denote the time point for a literal B_i as $t(B_i)$.

Since multiple neurons can fire at the same time, in the context of neuron diagrams there can be simultaneous events. In CP-logic, in each node only one law is allowed to be applied. Therefore it is not the case that each node in a branch of a probability tree corresponds to a separate time point, but rather sets of consecutive nodes – with variable size – do.

Concretely, a single diagram D corresponds to a set of branches B that all have the interpretation $\mathbf{v}$ in their leaf node. Each branch $b \in B$ contains the same edge-node pairs, but in different orders. The different orders of the application of the laws breaks down as follows. At each time point t, there may be several laws that are applied according to the temporal interpretation of the diagram D. (For example, $A \leftarrow$ and $C \leftarrow$.) If there are more than one, then there is one branch in B for each ordering of these laws. This reflects the fact that in a neuron diagram, all laws which have the same effect, symmetrically overdetermine this effect. Hence each branch can be divided into several subbranches, where each subbranch corresponds to one time point. The order of the subbranches is fixed, but the order in which laws are applied in each subbranch is completely open. As will become clear in the next chapter, in many cases the order of the laws does not matter and one can choose any $b \in B$ as an adequate translation of

the assignment $\mathbf{v}$.

Part I

A General Framework for Defining and Extending Actual Causation using CP-logic

Chapter 3

A General Definition of Actual Causation using CP-logic

This chapter was previously published together with Chapter 4 as:
Beckers, S., and Vennekens, J. (2016a). A general framework for defining and extending actual causation using cp-logic. *International Journal for Approximate Reasoning*, *77*, 105–126.

Further, parts of this chapter were previously published as:
Beckers, S., and Vennekens, J. (2015b). Towards a general framework for actual causation using cp-logic. In *Proceedings of the 2nd international workshop on probabilistic logic programming co-located with iclp* (Vol. 1413, pp. 19–38).

3.1 Introduction

In this chapter we will use CP-logic to formulate a general, parametrised, probabilistic definition of actual causation. The purpose of doing so is to obtain a simple and systematic method for comparing and constructing definitions of actual causation. We show how four existing definitions of actual causation can be reformulated and generalised as instantiations of our general definition.

First we present a reformulation of a previous definition developed by ourselves. This definition was orginally formulated in CP-logic, and already used some of the concepts defined in this chapter. In fact it can be seen as a precursor to our general definition.

Second we turn to work by Ned Hall. In his influential article "Two Concepts of Causation" (Hall, 2004), he argues for a view of causation as divided into two concepts (as the title suggests). The bottomline is that the type of examples in which we attribute causation is too diverse to be captured by a single concept. We reformulate and generalise his two concepts as instantiations of our general definition.

Comparing these three reformulated definitions then becomes straightforward. As a result, we see that our first definition is a compromise between the other two. This is confirmed by looking at the diverse examples discussed by Hall.

Our fourth definition is based on later work of Hall (2007). Unsatisfied with his earlier failure to incorporate all intuitive examples of causation into a single concept, he proposes a new definition as a compromise between the two concepts of causation. We reformulate and generalise this definition as an instantiation of our definition as well.

In the current chapter the focus lies on the flexibility offered by our general definition in expressing different definitions, rather than on the search for the right definition of causation. In the conclusion we briefly compare our general definition to similar work done in the structural equations literature.

Chapter 4 adds to our general definition an extension that allows it to take into account considerations of normality. The combination of both aspects, the general definition and its extension, is what makes up our general framework. In Chapter 5 we come back to the definitions here developed, and assess both their merits and the problems they face.

3.2 Defining Actual Causation Using CP-logic

Throughout the following chapters, we assume that we are given a CP-theory T and an actual story b in which both C and E occurred, and that we are interested in whether or not C caused E. We hereby limit ourselves to causation between literals C and E, but our account can be generalised to include complex causes and effects as well. By Con we denote the quadruple (T,b,C,E), and refer to this as a *context*.

3.2.1 Actual Causation in General

For reasons of simplicity, the majority of approaches (including Hall) only consider actual causation in a deterministic setting. Further, it is taken for granted that the actual values of all variables are given. In such a context, counterfactual dependence of the event E on C is expressed by the conditional: *if* $do(\neg C)$ *then* $\neg E$, where it is assumed that all exogenous variables take on their actual values. In our probabilistic setting, the latter translates into making those laws that were actually applied deterministic, in accordance with the choices made in the story. However, in many examples, the story does not specify the actual value of all exogenous variables. Looking back at our earlier Late Preemption example, if Suzy is prevented from throwing her rock, then we cannot say what the accuracy would have been had she done so. Hence, in a more general setting, it is required only that $do(\neg C)$ makes $\neg E$ possible. In other words, we get a probabilistic definition of counterfactual dependence:

Definition 2 (Dependence)**.** *E is* counterfactually dependent *on C in (T,b) if* $\mathbf{P}_{T^b}(\neg E \mid do(\neg C)) > 0$.

As counterfactual dependency lies at the heart of causation for all of the approaches we are considering, Dependence represents the most straightforward definition of actual causation.[1][2] It is however too crude and allows for many counterexamples, cases of preemption being the most famous.

More refined definitions agree with the general structure of the former, but modify the theory T in more subtle ways than T^b does. We identify two different kinds of laws in T, that should each be treated in a specific way.

The first are the laws that are *intrinsic* with respect to the given context. These are laws whose outcome is fixed, in the sense that in any counterfactual story we might consider, they will always produce the same outcome as they did in the actual story. Thus, intrinsic laws should be made deterministic in accordance with b.

The second are laws that are *irrelevant* in the given context. These are laws that played no part in the causal process that caused E, and that we should therefore not take into account when trying to find out if C was a cause of E or not. Thus, irrelevant laws should simply be ignored.

[1] This definition is similar in spirit to that of a *partial explanation* given in (Halpern & Pearl, 2005b). There the probability measures the *goodness* of the explanation, here it measures the *importance* of the cause.

[2] Fenton-Glynn (2015) uses these counterfactual probabilities to extend the definition of causation from Halpern and Pearl (2005a) to probabilistic structural equations. His focus lies with incorporating the idea that causes are "probability-raising" with regards to their effect, into a counterfactual account. In this manner, two traditional approaches to actual causation are combined. This project stands somewhat orthogonal to our current investigation. In future work it would be interesting to see how his results can be integrated into our framework.

Together, the methods of determining which laws are intrinsic and irrelevant, respectively, will be the parameters of our general definition. Suppose we are given two functions *Int* and *Irr*, which both map each context (T, b, C, E) to a subset of the theory T. With these, we define actual causation as follows:

Definition 3 (Actual causation given *Int* and *Irr*)**.** *Given the context Con, we define that C is an actual cause of E if and only if E is counterfactually dependent on C when replacing T^b with the theory T^* that we construct as:*

$$T^* = [T \setminus (Irr(Con) \cup Int(Con))] \cup Int(Con)^b.$$

For instance, the naive approach that identifies actual causation with counterfactual dependence corresponds to taking Irr as the constant function $\{\}$ and $Int(Con)$ as $\{r \in T \mid r \text{ was applied in } b\}$. From now on, we use the following, more legible notation for a particular instantiation of this definition:

Dependence-Irr. *No law r is irrelevant.*

Dependence-Intr. *A law r is intrinsic if r was applied in b.*

The following straightforward theorem expresses that this formulation of dependence is equivalent to the original in Definition 2.

Theorem 1. *Given the context Con, C is an actual cause of E given* Dependence-Irr *and* Dependence-Intr *iff E is counterfactually dependent on C.*

If desired, we can order different causes by their respective counterfactual probabilities $\mathbf{P}_{T^*}(\neg E \mid do(\neg C))$, as this indicates how important the cause was for E. Note however that Definition 2 reduces to a standard deterministic definition of counterfactual dependence if all CP-laws are deterministic. In that case, our general Definition 3 becomes deterministic as well.

3.3 Beckers and Vennekens 2012 Definition

A recent proposal for a definition of actual causation was originally formulated in (Vennekens, 2011), which we have slightly modified in (Beckers & Vennekens, 2012). Here, we summarize the basic ideas of the latter, and refer to it as *BV12*. We reformulate this definition in order to fit into our framework of Definition 3. It is easily verified that both versions are equivalent.

Because we want to follow the actual story as closely as possible, the condition for intrinsicness is exactly like before: we force all laws that were applied in b to have the same effect as they had in b.

To decide which laws were relevant for causing E in our story, we start from a simple temporal criterion: every law that was applied after the effect E took place is irrelevant, and every law that was applied before isn't. For example, to figure out why the bottle broke in the *Late Preemption* example, law (2.4) is considered irrelevant, because the bottle was already broken by the time Billy's rock arrived. For laws that were not applied in b, we distinguish laws that could still be applied when E occurred, from those that could not. The first are considered irrelevant, whereas the second aren't. This ensures that any story b' that is identical to b up to and including the occurrence of E provides the same judgements about the causes of E, since any law that is not applied in b but is applied in b', must obviously occur after E.

BV12-Irrelevant. *A law r is irrelevant if r was not applied before E in b, although it could have. (I.e., it was not impossible at the time when E occurred.)*

BV12-Intrinsic. *A law r is intrinsic if r was applied in b.*

The following theorem expresses that the current formulation is equivalent to the original definition.

Theorem 2. *Given the context Con, C is an actual cause of E given* BV12-Irr *and* BV12-Intr *iff C is an actual cause of E as defined in (Beckers & Vennekens, 2012).*

3.4 Hall 2004 Definitions

Hall (2004) claims that it is impossible to account for the wide variety of examples in which we intuitively judge there to be actual causation by using a single, all-encompassing definition. Therefore he defines two different concepts which both deserve to be called forms of causation but are nonetheless not co-extensive.

3.4.1 Dependence

The first of these is simply Dependence, as stated in Definition 2. As mentioned earlier, Hall only considers deterministic causal relations, and thus the probabilistic counterfactual will either be 1 or 0.

3.4.2 Production

The second concept tries to express the idea that to cause something is to bring it about, or to *produce* it. The original, rather technical, definition is discussed in the next

section, but the following informal version suffices to understand how it works: C is a producer of E if there is a directed path of firing neurons in the diagram from C to E. In our framework, this translates into the following.

Production-Irr. *A law r is irrelevant if r was not applied before E in b, or if its effect was already* **true** *when it was applied.*

Production-Intr. *A law r is intrinsic if r was applied in b.*

Theorem 3. *Given a neuron diagram D with corresponding equations M and assignment to its variables* **v**. *Consider the CP-logic theory T, and a story B, that we get when applying the translation from Section 2.4. C is a* producer *of E in the diagram according to Hall (2004) iff C is an actual cause of E given* Production-Irr *and* Production-Intr.

The CP-logic version of production offers a way to make sense of causation by omission. That is, just as with all of the definitions in our framework in fact, we can extend it to allow negative literals such as $\neg C$ to be causes as well.

In order to prove the above theorem, we first need to formulate Hall's definition of production.

3.4.3 Proof of Theorem 3

First we need to explain some terminology that Hall uses. On Hall's account, an *event* corresponds to the *firing of a neuron*, in other words, a variable taking on its deviant value **true**. A *structure* is a temporal sequence of sets of events that unfolds according to the structural equations of some neuron diagram and according to its temporal interpretation. A branch, or a sub-branch, would be the corresponding concept in CP-logic.

Two structures are said to *match intrinsically* when they are represented in an identical manner. The reason why Hall uses this term, is because even though we use the same variable for an event occurring in different circumstances, strictly speaking they are not the same. This is mainly an ontological issue, which need not detain us for our present purposes.

A set of events S is said to be *sufficient* for another event E, if the fact that E occurs follows from the causal laws, together with the premisse that S occurs at some time t, and no other events occur at this time. For example, in the diagrams from Figures 2.3 and 2.4, $\{A,C\}$ is one sufficient set for E, and $\{B\}$ is another one. The set $\{A\}$ is sufficient for both $\{B\}$ and $\{E\}$. A set is *minimally* sufficient if it is sufficient, and no proper subset is.

Now we can state the precise definition of production as it occurs in (Hall, 2004, p.25).

> We begin as before, by supposing that E occurs at t', and that t is an earlier time such that at each time between t and t', there is a unique minimally sufficient set for E. But now we add the requirement that whenever t_0 and t_1 are two such times ($t_0 < t_1$) and S_0 and S_1 the corresponding minimally sufficient sets, then
>
> - for each element of S_1, there is at t_0 a unique minimally sufficient set; and
> - the union of these minimally sufficient sets is S_0.
>
> ...
>
> Given some event E occurring at time t' and given some earlier time t, we will say that E has a *pure causal history* back to time t just in case there is, at every time between t and t', a unique minimally sufficient set for E, and the collection of these sets meets the two foregoing constraints. We will call the structure consisting of the members of these sets the "pure causal history" of E, back to time t. We will say that C is a *proximate cause* of E just in case C and E belong to some structure of events S for which there is at least one nomologically possible structure S' such that (i) S' intrinsically matches S; and (ii) S' consists of an E-duplicate, together with a pure causal history of this E-duplicate back to some earlier time. (In easy cases, S will itself be the needed duplicate structure.) Production, finally, is defined as the ancestral [i.e., the transitive closure] of proximate causation.

Note that Hall's definition only applies to events as candidate cause or effect. Hence we only need to prove equivalence between the two definitions in case both C and E are positive literals. Our definition is a generalisation to all literals. We should point out that it judges negative literals not to have any causes, so the only generalisation of interest is that of the causes.

Proof. Assume we have a neuron diagram D, with assignment $\mathbf{v}$. Say T is the CP-logic theory that is the translation of the equations of the diagram, and B is the set of branches representing the story. The different branches in B are all made up of the same edges and nodes, except that they occur in a different order. Recall that C is an actual cause of E in a set of branches iff C is an actual cause of E in any of them.

By *Production-Intr*, regardless of which $b \in B$ we use to construct the theory T^*, it will consist of the same deterministic laws. Concretely, all its laws are of the form: $V_i \leftarrow$, or $V_j \leftarrow B_1 \wedge \ldots \wedge B_m$, where the number of atoms in the conjunction is at least one. (The presence of at least one atom is due to the fact that a neuron diagram does not allow for direct causation by omission, as we noted before.)

Therefore any probability tree for T^* consists of a single branch, determining a unique assignment for all the variables, namely the assignment **v** that holds in the leaf of each $b \in B$. By *Production-Irr*, for each variable V_i that is assigned **true** in **v**, for each choice of $b \in B$ there is exactly one law in T^* which has V_i in its head. For each $b \in B$, we denote by $Pos_b(V_i)$ the set of all positive literals that occur in the body of this unique law.

We will prove the equivalence by induction over $t(E)$.

Base case: $t(E) = 0$.

This implies that E is determined directly by the exogenous variables. So for each $b \in B$ the only law containing E in its head in T^* is $E \leftarrow$. Clearly, by both definitions there are no producers of E in this case.

Induction case:

Assume that the equivalence holds for any effect V_i such that $t(V_i) \leq n$. We need to prove that it holds as well for E if $t(E) = n+1$.

For each $b \in B$, clearly $Pos_b(E)$ is a minimally sufficient set for E at time n. If we consider the structure consisting of $Pos_b(E) \cup \{E\}$, then $S_0 = Pos_b(E)$ is a unique minimally sufficient set for $S_1 = E$ at time n. Therefore $Pos_b(E) \cup \{E\}$ is a pure causal history of E back to time n. Therefore for each $b \in B$, each $V_j \in Pos_b(E)$ is a proximate cause of E, and thus also a producer of E according to Hall's definition. Further, it is easy to see that for each $b \in B$ and $V_j \in Pos_b(E)$, $\mathbf{P}_{T^*}(\neg E | do(\neg V_j)) = 1$. Hence it is also a producer according to our definition.

Now consider a variable V_j that is assigned **true** by **v**, and such that for all $b \in B$, $V_j \notin Pos_b(E)$. Then there is no minimally sufficient set for E that contains V_j, so it is not a producer of E according to Hall's definition. Also, it is easy to see that for each $b \in B$, we have $\mathbf{P}_{T^*}(\neg E | do(\neg V_j)) = 0$. Hence it is not a producer according to our definition either.

Therefore the equivalence holds for all producers of E at time n.

Remains to be shown that it holds for all producers of E at earlier times. We start with the implication from left to right.

Assume V_j is a producer of E at some time $m < n$, according to Hall's definition. This means there exists a V_k such that $t(V_k) = n$, V_j is a producer of V_k, and V_k is a producer of E, according to Hall's definition. By the above, this implies that there is at least one $b_i \in B$ so that $V_k \in Pos_{b_i}(E)$. As above, for any such b_i it holds that $\mathbf{P}_{T^*}(\neg E | do(\neg V_k)) = 0$ for the theory T^* constructed using b_i.

By the induction hypothesis, we get that V_j is also a producer of V_k according to our definition. Hence there exists a branch $b_2 \in B$ so that $\mathbf{P}_{T^*}(\neg V_k | do(\neg V_j)) = 0$ for the

theory T^* constructed using b_2. If $V_k \in Pos_{b_2}(E)$, then as above $\mathbf{P}_{T^*}(\neg E|do(\neg V_k)) = 1$.

Assume that for all of the b_i above, it holds that $Pos_{b_2}(E) \neq Pos_{b_i}(E)$. This means that in each $b \in B$, there are at least two laws such that its effect was E. Denote by r_1 the first law which had as its effect E in b_2, and let r_2 denote the first law which had as its effect E in any of the b_i. Thus $Pos_{b_2}(E) \neq Pos_{b_i}(E)$ only because in b_2 law r_1 was applied before r_2. Given the different orders that exist in B as explained above, this implies that there also exists another branch $b_3 \in B$ so that $\mathbf{P}_{T^*}(\neg V_k|do(\neg V_j)) = 1$ for the theory T^* constructed using b_3, and also $V_k \in Pos_{b_3}(E)$, implying that $\mathbf{P}_{T^*}(\neg E|do(\neg V_k)) = 0$. To see why, we have a look at the set B.

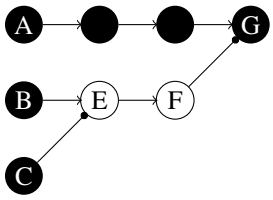
We have arrived at a branch b_3 so that both $\mathbf{P}_{T^*}(\neg V_k|do(\neg V_j)) = 1$ and $\mathbf{P}_{T^*}(\neg E|do(\neg V_k)) = 1$. Given the structure of T^*, namely the fact that it is deterministic, and that for each **true** variable V_i in the assignment $\mathbf{v}$ there is exactly one law with V_i in the head, it follows that $\mathbf{P}_{T^*}(\neg E|do(\neg V_j)) = 1$. Therefore V_j is a producer of E according to our definition as well.

Now we prove the reverse implication.

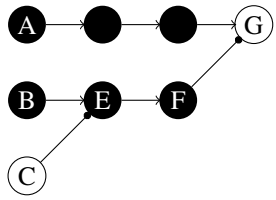
Assume that V_j is a producer of E according to our definition, such that $t(V_j) = m < n$. So there is a branch b such that $\mathbf{P}_{T^*}(\neg E|do(\neg V_j)) = 1$ for the theory T^* constructed using b. Since we have $\neg E$ in $T^*|do(\neg V_j)$, there must be some $V_k \in Pos_b(E)$ such that $\mathbf{P}_{T^*}(\neg V_k|do(\neg V_j)) = 1$. By the induction hypothesis, V_j is a producer of V_k according to Hall's definition. Also, $\mathbf{P}_{T^*}(\neg E|do(\neg V_k)) = 1$, and hence by the above equivalence regarding producers of E at time n, we know that V_k is a producer of E according to Hall's definition. Since by Hall's definition production is transitive, the result follows. □

3.5 Illustration: Double Prevention

To illustrate the application of the three definitions presented so far, we have a look at *Double Prevention*, an example from Hall (2004).



(**c** fires)



(if **c** hadn't fired)

The term *Double Prevention* refers to the fact that C prevents F from preventing G, which leads most people to intuitively consider C a cause of G. Hall (2004) presents this example as a challenge to Production, since as we are about to see, C does not produce G. There is Dependence between G and C, which illustrates Hall's point that one cannot get by with only Production.

For the CP-logic translation of this example we make us of the following theory on the left and one of its probability trees on the right, where p and q represent some probabilities. (We leave out the labels indicating which laws are applied, as the edges already contain that information in this case.)

Theory T

$(A : p) \leftarrow .$

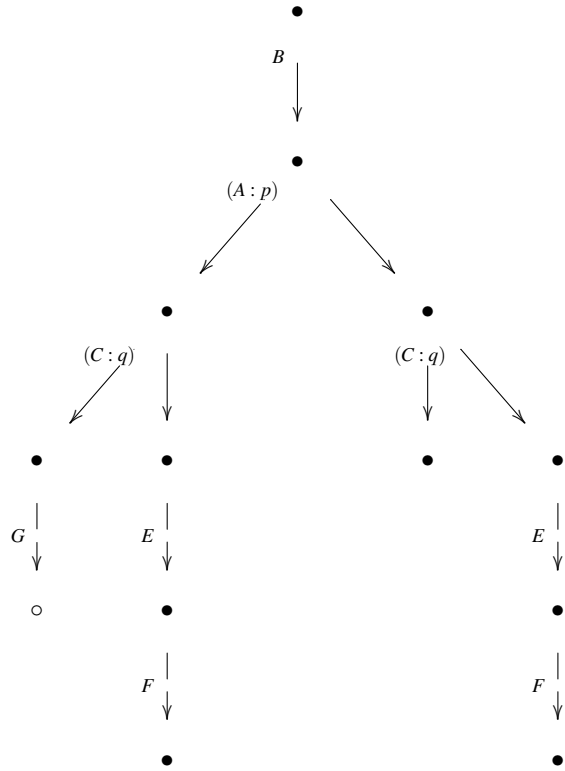
$B \leftarrow .$

$(C : q) \leftarrow .$

$E \leftarrow B \wedge \neg C$

$F \leftarrow E.$

$G \leftarrow A \wedge \neg F.$



The *Double Prevention* example then corresponds to the leftmost branch of the probability tree. (We focus on a single branch to represent the diagram, rather than a set of branches, since the outcome will be the same for each branch.) As a first step we focus on *Intrinsicness*. All three of Dependence, Production and the BV12 definition, share the same intrinsicness condition, namely that for a given story b, the theory T should be updated to the theory T^b. Further, we need to apply the intervention $do(\neg C)$,

giving $T^b|do(\neg C)$.

Theory T^b

$A \leftarrow .$

$B \leftarrow .$

$C \leftarrow .$

$E \leftarrow B \wedge \neg C$

$F \leftarrow E.$

$G \leftarrow A \wedge \neg F.$

Theory $T^b|do(\neg C)$

$$A \leftarrow . \tag{3.1}$$

$$B \leftarrow . \tag{3.2}$$

$$E \leftarrow B \wedge \neg C. \tag{3.3}$$

$$F \leftarrow E. \tag{3.4}$$

$$G \leftarrow A \wedge \neg F. \tag{3.5}$$

Since Dependence considers all laws to be relevant, it takes T^b to be the modified theory T^*. Applying the definition of actual causation in this case gives $\mathbf{P}_{T^*}(\neg G|do(\neg C)) = 1$, so that Dependence judges C to be a full cause of G. The BV12 definition arrives at exactly the same conclusion: even though the laws (3.3) and (3.4) did not happen in the Double Prevention story, they were impossible, and are therefore considered relevant. For Production, it is a different matter, as it considers the laws (3.3) and (3.4) irrelevant, and therefore takes T^* to be the following theory:

Theory $T^*|do(\neg C)$

$A \leftarrow .$

$B \leftarrow .$

$G \leftarrow A \wedge \neg F.$

In this case $\mathbf{P}_{T^*}(\neg G|do(\neg C)) = 0$, meaning that according to the Production account C wasn't a cause of G.

3.6 A Compromise between Dependence and Production

Hall's motivation for introducing two concepts of causation is the existence of examples which exhibit two different types of relations that we both call causation. Therefore any account of causation in terms of a single concept which also wishes to accommodate the examples Hall discusses, can be expected to consist in some form of mixture of the two relations.

When we compare the previous three definitions, we see that they agree on the intrinsicness condition. The difference between them is to be found only in the strength of the irrelevance condition. Production has the weakest condition, dismissing two types of laws as irrelevant: those that were not active before E, and also those whose effect was already **true**. CP-logic only dismisses a subset of the first type: only those that were not active before E but could have been. Dependence, finally, has the strongest irrelevance condition, as it does not dismiss any law. The BV12 definition therefore naturally presents itself as a compromize between Hall's two definitions.

This perspective is further confirmed when we take a look at the problematic examples Hall discusses. In Table 3.1 we summarize the verdict of the different definitions regarding these examples. Not only would a detailed discussion of all of these lead us too far, it would provide little added value, as the BV12 definition agrees with Hall's judgments on all of the examples except for his version of *Early Preemption* (EP), which will be discussed in depth in Part II.

The first column states the examples we refer to, where the figure numbers are those found in (Hall, 2004), with the exception of the last three examples which come from (Hall, 2007). EP stands for *Early Preemption*, LP stands for *Late Preemption*, DP stands for *Double Prevention*, Om stands for *Omission*, N-ET stands for *Non-Existent Threats* and S-C stands for *Short-Circuit*. (For completeness we point out that the model Hall uses for LP is problematic. Part II contains a detailed discussion on the problems that arise when trying to model LP using structural equations.)

The last column contains Hall's intuitions about the examples. (The 1 and 0 stand for the judgment that in this example there is actual causation regarding the variables under discussion or not, and a question mark indicates that the judgment is unclear.) The second column contains the answers given by the most well-known definition of actual causation, that of Halpern and Pearl (2005a) which we presented in Section 2.2.1. We add this simply for a quick comparison, without commenting on it. We discuss this definition in Part II. We have also added Hall's more recent definition from Hall (2007), which we will discuss next.

The main lessons to be learned from this table are the following:

- Unsurprisingly, Hall's 2007 definition accords with his earlier intuitions on all examples.
- Except for EP, the BV12 definition is also able to analyze all the examples in accordance with Hall's intuitions without seeking recourse in two concepts of causation.
- The examples in Fig 14 and 15 refer to what Hall called his 'unfinished business', because they exhibit a relation of causation which can not be described

Table 3.1: Comparison of definitions

Example	HP	Dependence	Production	BV12	Hall2007	Hall's intuition
EP	1	0	1	0	1	1
LP	1	0	1	1	1	1
DP	1	1	0	1	1	?
Fig 5	1	1	0	1	1	1
Om	1	1	0	1	0	?
Fig 12	1	0	1	0	0	?
Fig 14	1	0	0	1	1	1
Fig 15	1	1	?	1	1	1
Switch	1	0	1	0	0	0
N-ET	1	0	0	0	0	0
S-C	1	0	0	0	0	0

appropriately in terms of Dependence and Production. Both BV12 and Hall 2007 are able to handle them.

- The last three examples are presented as counterexamples to the HP-definition, in that it provides counterintuitive responses for them. Both BV12 and Hall 2007 do not suffer from this. The *Switch* example will be discussed in Chapter 5 and Part II.

In the current section our intent was twofold: firstly, to illustrate the usefulness of expressing Hall's two concepts of causation in terms of our general definition by comparing it with BV12, and, secondly, to show that the BV12 definition succeeds in accepting all but one of Hall's intuitions regarding examples of causation whilst using a single concept of causation. In the next section we will discuss Hall's most recent definition, which enriches both aspects of the current discussion. First, by expressing yet another definition in CP-logic – using the same terminology of irrelevance and intrinsicness – the generality of our approach is further confirmed. Second, as the new definition also provides an adequate solution to the problems with Hall's earlier view, a further comparison between it and the BV12 definition is called for. This comparison will be the subject matter of Chapter 5.

3.7 Hall 2007

One of the currently most refined concepts of actual causation is that of Hall (2007). Although Hall uses structural equations as a practical tool, he is of the opinion that

intuitions about actual causation are best illustrated using neuron diagrams. A key advantage of these diagrams, which they share with CP-logic, is that they distinguish between the default and deviant state of a variable.

Defenders of the structural equations approach – as could be expected – have in return critiziced Hall's account. Hitchcock (2009) comes up with several examples which cannot be handled properly by Hall's definition, which we will discuss in Chapter 5. But part of his criticism is also focused on Hall's choice of neuron diagrams as models for all types of causal mechanisms (Hitchcock, 2009, p. 398):

> ...neuron diagrams are poor heuristic devices for suggesting possible causal structures. Neuron diagrams privilege one particular pattern of counterfactual dependence – a neuron fires if it is stimulated by at least one other neuron, and inhibited by none. Alternative patterns of dependence can be represented on an ad hoc basis;... But neuron diagrams do nothing to actively suggest such alternative patterns of dependence. Structural equations, by contrast, represent patterns of dependence algebraically or truth-functionally, which greatly facilitates the enumeration of possible patterns of dependence. Thus even if Hall is correct that structural equation models are merely representational tools, I think he underestimates the importance of having a suitable representation.

A neuron diagram, and thus Hall's approach as well, is very limited in the kind of examples it can express. In particular, neuron diagrams can only express deterministic causal relations, and they lack the ability to directly express *causation by omission*, i.e., that the absence of C by itself causes E, as in the law $E \leftarrow \neg C$. Hall's solution is to argue against causation by omission altogether. By contrast, we will offer an improvement of Hall's account that generalizes to a probabilistic context, and can also handle causation by omission. In short, we propose CP-logic as a way of overcoming the shortcomings of both structural equations and neuron diagrams.

The idea behind Hall's definition is to check for counterfactual dependence in situations which are reductions of the actual situation, where a reduction is understood as "a variant of this situation in which *strictly fewer* events occur". In other words, because the counterfactual dependence of E on C can be masked by the occurrence of events which are extrinsic to the actual causal process, we look at all possible scenario's in which there are less of these extrinsic events. Hall puts it like this (2007)[p. 129]:

> Suppose we have a causal model for some situation. The model consists of some equations, plus a specification of the actual values of the variables. Those values tell us how the situation *actually* unfolds. But the same system of equations can also represent *nomologically possible variants*:

> just change the values of one or more exogenous variables, and update the rest in accordance with the equations. A good model will thus be able to represent a range of variations on the actual situation. Some of these variations will be – or more accurately, will be modeled as – *reductions* of the actual situation, in that every variable will either have its actual value or its default value. Suppose the model has variables for events C and E. Consider the conditional
>
> **if $C = 0$; then $E = 0$**
>
> This conditional may be true; if so, C is a cause of E. Suppose instead that it is false. Then C is a cause of E iff there is a reduction of the actual situation according to which C and E still occur, and in which this conditional is true.

Rather than speaking of fewer events occuring, in this definition Hall characterizes a reduction in terms of whether or not variables retain their actual value. This is because in the context of neuron diagrams, an event is the firing of a neuron, which is represented by a variable taking on its deviant value, i.e., the variable *becoming true*. In the dynamic context of CP-logic, this corresponds to the transition in a probability tree (i.e., the application of a causal law) that makes such a variable true. Therefore we take a reduction to mean that no law is applied such that it makes a variable true that did not become true in the actual setting.

To make this more precise, we introduce some new formal terminology. Let d be a branch of a probability tree of the theory T. $Laws_d$ denotes the set of all laws that were applied in d. The resulting effect of the application of a law $r \in Laws_d$ – i.e., the disjunct of the head which was chosen – will be denoted by r_d, or by 0 if an empty disjunct was chosen. The set of true variables in the leaf of d will be denoted by $Leaf_d$. Note that $\forall d : Leaf_d = \bigcup_{r \in Laws_d} r_d$.

A branch d is a *reduction* of b if $\forall r \in Laws_d : r_d = 0 \vee \exists s \in Laws_b : r_d = s_b$. Or, equivalently, $Leaf_d \subseteq Leaf_b$.

A reduction of b in which both C and E occur – i.e., hold in its leaf – will be called a (C,E)-reduction. The set of all of these will be denoted by $Red_b^{(C,E)}$. These are precisely the branches which are relevant for Hall's definition.

Definition 4 (Hall's definition in CP-logic)**.** *We define that C is an* actual cause *of E if* $(\exists d \in Red_b^{(C,E)} : \mathbf{P}_{T^d}(\neg E | do(\neg C)) > 0)$.

Theorem 4 shows the correctness of our translation. We point out that none of the definitions in this section focus on the order in which laws are applied in a branch.

Hence we can translate the assignment $\mathbf{v}$ of a neuron diagram into a single branch b, rather than into a set of branches B, without loss of generality.

Theorem 4. *Given a neuron diagram D with its corresponding equations M, and an assignment to its variables $\mathbf{v}$. Consider the CP-logic theory T and story b that we get when applying the translation from Section 2.4. Then C is an actual cause of E in the diagram according to Hall's definition quoted above iff C is an actual cause of E in b and T according to Definition 4.*

To facilitate the proof of this theorem, we introduce the following lemma.

Lemma 1. *Given a neuron diagram D with corresponding equations M and assignment to its variables $\mathbf{v}$, and $\{C,E\} \subseteq \mathbf{v}$. Consider the CP-logic theory T, and a story b, that we get when applying the translation from Section 2.4. Then a neuron diagram R is a reduction of D in which both C and E occur iff its translation d – another branch of T – is a (C,E)-reduction of b.*

Proof. Say $\mathbf{u}$ is the context for D, i.e., $(M,\mathbf{u})$ gives the assignment $\mathbf{v}$. A reduction of D in which both C and E occur is the result of changing some of the exogenous variables from **true** to **false**, changing $\mathbf{u}$ into $\mathbf{u}'$, in such a way that $(M,\mathbf{u}')$ gives an assignment $\mathbf{v}'$ so that the set of true variables in $\mathbf{v}'$ is a subset of the set of true variables in $\mathbf{v}$.

Recall that each u_i determines a single endogenous variable, say A, and that the translation of this into CP-logic is the law $(A : *) \leftarrow$. A change from u_i into $\neg u_i$ is thus translated into selecting the empty disjunct 0 as a child-node, instead of selecting A, in the node where the law $(A : *) \leftarrow$ is applied. Therefore the change from $\mathbf{u}$ into $\mathbf{u}'$ translates simply the change from b into a different branch d of the same probability tree of T. The same holds in the other direction: changing b into a branch d by selecting different nodes so that more empty disjuncts are chosen when applying the non-deterministic laws, translates into changing $\mathbf{u}$ into $\mathbf{u}'$ so that some true exongeous variables become false. Since the equations in M and the CP-theory T determine identical assignments given a context u, or given its translation into a choice of nodes in CP-logic, and the assignment fully determines whether or not something is a reduction containing C and E, the equivalance follows. □

Now we prove Theorem 4.

Proof. The conditional $C = 0$ in the statement "if $C = 0$; then $E = 0$", made relative to the causal setting $(M,\mathbf{u})$, is to be interpreted as a counterfactual locution, i.e., it means that $(M_{do(\neg C)},\mathbf{u}) \models \neg E$. Since for any story d the theory T^d is deterministic, this is equivalent to $\mathbf{P}_{T^d}(\neg E|do(\neg C)) = 1$ for any branch d that is a CP-logic translation of $(M,\mathbf{u})$. Given the equivalence between (C,E)-reductions in neuron diagrams and CP-logic from the previous lemma, the result follows. □

At first sight, Definition 4 does not fit into the general framework we introduced earlier, because of the quantifier over different branches. However, we will now show that for a significant group of cases it actually suffices to consider just a single T^*, which can be described in terms of irrelevant and intrinsic laws.

Rather than looking at all of the reductions separately, we single out a minimal structure which contains the essence of our story. This structure will be based on the set of all laws that are *necessary* for a reduction, in the sense that they are applied in each reduction and that, moreover, they are always applied in the same way, i.e., with the same outcome.

Definition 5. *A law r is* necessary *for a story b if*

- $\forall d \in Red_b^{(C,E)} : r \in Laws_d$ *and*
- $\forall d, e \in Red_b^{(C,E)} : r_d = r_e$.

We define $Nec(b)$ as the set of all necessary laws for b.

In general it might be that there are two (or more) laws which are unnecessary by themselves, but at least one of them has to be applied as it was in b. Consider for example the following CP-theory.

$$C \leftarrow . \qquad (A : p) \leftarrow C. \qquad E \leftarrow A.$$
$$(B : q) \leftarrow C. \qquad E \leftarrow B.$$

In the story where C causes both A and B, each of those being sufficient for E, neither the second nor the third law is necessary for E. Yet it is clear that at least one of them has to be applied to get E. In cases where this complication does not arise, we shall say that the story is *simple*.

First, some helpful terminology: an r-variant of a branch b is any branch b' which coincides with b up to the application of r, but which then selects a different disjunct $r_{b'} \neq r_b$ from the head of r. (Where possibly $r_b = 0$ or $r_{b'} = 0$.) Note that if r is deterministic, there are no r-variants.

Definition 6. *A story b is* simple *if the following holds:*

$\forall d \in Red_b^{(C,E)}$, for all non-deterministic $r \in Laws_d \setminus Nec(b)$, and for each disjunct r_s of r: $\exists e \in Red_b^{(C,E)}$ so that e is an r-variant of d with $r_e = r_s$.

Informally, a story is simple if there are no limitations on the choices made for laws that are not necessary.

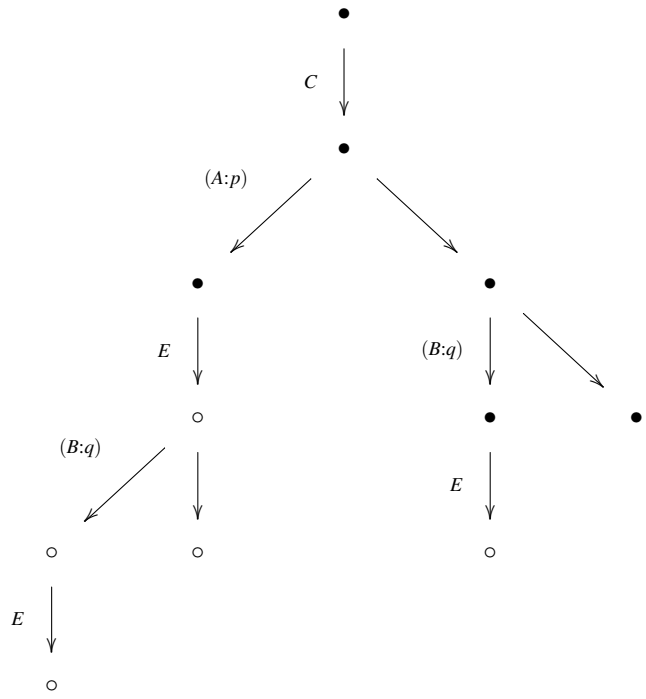


Figure 3.1

We illustrate this concept by looking at a probability tree for the previous example in Figure 4.1, to show that the previous story is not simple. The left-most branch is a formal counterpart of this story. Except for the right-most branch in which E is **false**, all branches are (C,E)-reductions of b. To see that the second law r is not necessary, observe that for any branch d in the left-side of the tree, $r_d = A$, whereas for any branch e in the right-side, $r_e = 0$. Similarly, the third law is not necessary either.

Now consider the second branch from the right, d, and the third law, r', that results in B in the fourth node: there does not exist a (C,E)-reduction e of b that is identical to d up to the third node (i.e., up to the application of r'), but different regarding the fourth node (i.e., $B = r'_d \neq r'_e$). Hence, b is not simple.

We are now in a position to formulate a theorem that will allow us to adjust Hall's definition into our framework.

Theorem 5. *If* $(\exists d \in Red_b^{(C,E)} : \mathbf{P}_{T^d}(\neg E|do(\neg C)) > 0)$*, then it holds that* $\mathbf{P}_{T^{Nec(b)}}(\neg E|do(\neg C)) > 0$*. If b is simple, then the reverse implication holds as well.*

Proof. We start by comparing T^d and $T^{Nec(b)}$. Both are constructed out of T, by making some laws deterministic. By definition, any law $r \in Nec(b)$ is made deterministic in both T^d and $T^{Nec(b)}$, with the outcome r_b. The only difference between both theories is that there may also be other non-deterministic laws from T that are made deterministic in T^d, contrary to $T^{Nec(b)}$, where these laws remain non-deterministic. Therefore there are more possible stories, or branches, according to $T^{Nec(b)}$. Specifically, every branch e that occurs in a probability tree of T^d also occurs in a probability tree of $T^{Nec(b)}$. The only difference between both branches might be the probability by which it occurs, but obviously in both cases it will be strictly positive. (The latter is simply a property of any branch in CP-logic.)

With this insight, proving the first implication is easy. Assume we have a $d \in Red_b^{(C,E)}$ such that $\mathbf{P}_{T^d}(\neg E|do(\neg C)) > 0$. This implies that there is at least one branch e of a probability tree of $T^d|do(\neg C)$ for which $\neg E$ holds in its leaf. By the above explanation, it follows immediately that there is a branch e' of a probability tree of $T^{Nec(b)}|do(\neg C)$ with the same property, which is precisely what we had to prove.

Now we prove that if b is simple, the reverse implication holds as well. Assume there is a branch f of a probability tree of $T^{Nec(b)}|do(\neg C)$ for which $\neg E$ holds in its leaf, and b is simple. We need to proof that there exists a (C,E)-reduction d for which there exists a branch with the same property.

Let $Unn(b) = (T \setminus Nec(b))|do(\neg C)$, i.e., $Unn(b)$ are the non-deterministic laws in $T^{Nec(b)}|do(\neg C)$. (The "Unnecessary" laws, so to speak.) Clearly, the difference between any $T^d|do(\neg C)$ and $T^{Nec(b)}|do(\neg C)$ can consist only in the fact that some of the laws in $Unn(b)$ are deterministic in the former.

Let d be any (C,E)-reduction of b. Say n is the first node in f so that a law r is applied for which we cannot find a branch f' in a probability tree of $T^d|do(\neg C)$ such that $r'_f = r_f$. We need to show that there is another (C,E)-reduction of b such that it does allow for a branch identical to f up until $n+1$. If we do that, then by induction this applies to the entire branch f and we have the desired result.

Clearly, $r \in Unn(b)$. First, consider the possibility that $r \notin Laws_d$. This means that r appears in T^d in its original non-deterministic form, just as it does in $T^{Nec(b)}$. Hence it takes on the same form in both $T^{Nec(b)}|do(\neg C)$ and $T^d|do(\neg C)$. But this means that the choice r_f must be available in f' as well, contradicting our assumption. Hence it must be the case that $r \in Laws_d$.

Together with the fact that $r \in Unn(b)$, we get that $r \in Laws_d \setminus Nec(b)$. Since b is simple, this implies that $\exists e \in Red_b^{(C,E)}$ so that e is an r-variant of d, with $r_e = r_f$. Therefore e has the desired property that it allows for a branch identical to f up until $n+1$. This concludes the proof. □

Although the reverse implication is limited to simple stories, we do not consider this a severe restriction: all of the examples Hall discusses are simple, as are all of the classical examples discussed in the literature, such as *Early* and *Late Preemption*, *Symmetric Overdetermination*, *Switches*, etc.

As a result of this theorem, rather than having to look at all (C,E)-reductions and calculate their associated probabilities, we need only find all the necessary laws and calculate a single probability. If the story b is simple, then this probability represents an extension of Hall's definition, since they are equivalent if one ignores the value of the probability but for it being 0 or not. To obtain a workable definition of actual causation, we present a more constructive description of necessary laws.

Theorem 6. *If b is simple, then a law r that was applied in b is necessary iff none of the r-variants of b is a (C,E)-reduction.*

Proof. We denote the node in b in which r is applied by n. Thus $n+1$ is the child of n representing the selection of r_b, and the r-variants of b are all branches not passing through $n+1$.

We start with the implication from left to right, so we assume r is necessary. Assume $r_b = A$, then there is no $d \in Red_b^{(C,E)}$ for which $r_d \neq A$, hence there is no (C,E)-reduction which does not pass through $n+1$.

Remains the implication from right to left. Assume we have a law r such that all (C,E)-reductions of b in the tree to which b belongs pass through $n+1$. We proceed with a reductio ad absurdum, so we assume r is not necessary.

Trivially, b is a (C,E)-reduction of itself. Also, $r \in Laws_b \setminus Nec(b)$. Hence, by b's simplicity, there is a (C,E)-reduction e which is identical to b up to the application of r, but for which $r_e \neq r_b$. Thus e does not pass through $n+1$, contradicting our earlier assumption. This concludes the proof. □

With this result, we can finally formulate our version of Hall's definition, which we will refer to as Hall07.

Hall07-Irrelevant. *No law r is irrelevant.*

Hall07-Intrinsic. *A law r is intrinsic if r was applied in b, and none of the r-variants d of b is such that $\{C,E\} \subseteq \text{Leaf}_d \subseteq \text{Leaf}_b$.*

3.8 Comparison

Table 3.2 presents a schematic overview of the four definitions we have discussed. The columns and rows give the criteria for a law r of T to be considered intrinsic,

respectively irrelevant, in relation to a story b, and an event E. By $r \leq_b E$, we denote that r was applied in b before E occurred.

Table 3.2: Spectrum of definitions

Irrelevant	Intrinsic	
	$r \in Laws_b$	$r \in Nec(b)$
$\emptyset$	Dependence	Hall07
$\exists d : (d = b \text{ up to } E) \wedge r \geq_d E$	BV12	
$r \not<_b E \vee r_b <_b r$	Production	

Looking at this table, we can informally characterise the different definitions by describing which events are allowed to happen in the counterfactual worlds they take into consideration to judge causation:

- **Production**: Only those events – i.e., applications of laws making a variable **true** – which happened before E, and not differently – i.e., with the same outcome as in the actual story.
- **BV12**: Those events which happened before E, and not differently, and also those events which were prevented from happening by these.
- **Hall07**: All events, as long as those events that were necessary to E do not happen differently.
- **Dependence**: All events, as long as those events that did actually happen do not happen differently.

In order to illustrate the working of the definitions and their differences, we present an example:

Example 3. *Assassin decides to poison the meal of a victim, who subsequently Dies right before dessert. However, Murderer decided to murder the victim as well, so he poisoned the dessert. If Assassin had failed to do his job, then Backup probably would have done so all the same.*

The causal laws that form the background to this story are give by the following theory:

$$(Assassin : p) \leftarrow . \tag{3.6}$$

$$(Murderer : q) \leftarrow . \tag{3.7}$$

$$(Backup : r) \leftarrow \neg Assassin. \tag{3.8}$$

$$Dies \leftarrow Assassin. \tag{3.9}$$

$$Dies \leftarrow Backup. \tag{3.10}$$

$$Dies \leftarrow Murderer. \tag{3.11}$$

In this story, did *Assassin* cause *Dies*?

In order to apply the Hall07 definition, we first show that the story is *simple*, in the sense of Definition 7. Here $C = Assassin$, $E = Dies$ and b is the actual story. We denote law (3.6) by ^{1} and law (3.7) by r^2. Note that Hall07 does not refer to the order of a story at any point, nor do any of the other concepts which we defined in Section 3.7. Hence we may consider all stories in which all laws have the same outcome to be identical, regardless of the order in which the laws were applied.

Assassin has to be **true** in any (C,E)-reduction, which implies that r^1 will have *Assassin* as its chosen disjunct, i.e., $\forall d \in Red^{(C,E)} : r^1_b = Assassin$. By definition 5, this means that r^1 is necessary for our story. Since by law (3.9) *Assassin* guarantees *Dies*, it is easy to see that r^2 is not necessary for our story. Therefore, ignoring the order, there is only one (C,E)-reduction aside from the actual story itself: the story s for which $r^1 = Assassin$, and $r^2_s = 0$. Since the stories b and s are r^2-variants of each other, the criterion for simplicity is met.

We leave it to the reader to verify that in this case the left intrinsicness condition from the table applies to laws (3.6) and (3.7), whereas the right one only applies to (3.6). (Note that there is no use in checking whether the laws (3.9) to (3.11) are intrinsic or not, since these are already deterministic and hence it does not matter.) The second irrelevance condition only applies to law (3.11), whereas the third one applies to laws (3.8), (3.10) and (3.11). This results in the following probabilities representing the causal status of *Assassin*:

Production	BV12	Hall07	Dependence
1	$1-r$	$(1-r)*(1-q)$	0

Hence *Assassin* is a full cause according to Production, not a cause at all according to Dependence, and somewhere in between these two extremes according to the other two definitions.

Intuitively, most people would judge *Assassin* to be fully responsible for causing victim's death. Hence this particular example seems to speak in favour of Production. However, note that this example is clearly set in a normative context, since murdering people is – in almost all cases – judged to be wrong. One can easily come up with morally neutral examples using these CP-laws and the same story such that our intuitions would be different, for instance the following story:

Example 4. *Billy has set the alarm for six o'clock, at which time it goes off, so that he and Suzy make it in time to school. However, Suzy had put her alarm for five past six, which would have also left ample amount of time. If Billy had failed to put his alarm, then Mother probably would have done so all the same.*

In this story, it sounds quite reasonable to say that *Billy* is not a full cause of *Billy* and *Suzy* making it to school on time. We deliberately first chose an example that contains normative elements, because it is a general feature of all existing definitions of causation that they fail to do justice to such context-dependency. The next chapter extends our general definition of actual causation so that it can incorporate judgments of what is considered normal in a particular context.

3.9 Conclusion and Related Work

In this chapter we have used the formal language of CP-logic to formulate a general definition of actual causation, which we used to express four specific definitions: a previous proposal of our own, and three definitions based on the work of Hall. By moving from the deterministic context of neuron diagrams to the non-deterministic context of CP-logic, the latter definitions improve on the original ones in two ways: they can deal with a wider class of examples, and they allow for a graded judgment of actual causation in the form of a conditional probability. Also, comparison between the definitions is facilitated by presenting them as various ways of filling in two central concepts.

As a result, we found that both the BV12 and Hall07 definitions can be viewed as a compromise between two distinct and important concepts, Production and Dependence. Both of these concepts capture fundamental intuitions regarding causation, which will return in Part II.

Many other definitions exist in the counterfactual tradition. Rather than arguing for or against a particular definition, our aim in the current chapter was to develop a general parametrised definition as a tool for constructing, comparing, and modifying different definitions in a systematic way. We briefly discuss some other definitions in order to understand the relation of our work to that of others. A more detailed discussion of these definitions appears in Part II.

The most influential definition of actual causation to date is the *HP definition* of Halpern and Pearl (2005a). This definition is expressed using structural equations as they are developed by Pearl (2000). Despite – or because of – its popularity, it has been subjected to much criticism. Restricting ourselves to authors working within the counterfactual tradition in the spirit of Lewis (1973), we can divide the criticism into two types.

The first type criticises not just the HP definition itself, but also its formulation in terms of structural equations. We already mentioned Hall (2007) as a prominent example. In previous work we have also criticised the HP definition and structural equations in general (Beckers & Vennekens, 2012; Vennekens, 2011). The current chapter is a continuation of both lines of work. While structural equations are useful for a variety of purposes, we feel they lack certain key features when it comes to actual causation, which are present in CP-logic: true non-determinism in the endogenous part of the model, the distinction between default and deviant values, and a temporal semantics. It is possible to extend the language of structural models in various ways to incorporate such features (see, e.g., (Fenton-Glynn, 2015; Halpern & Pearl, 2005a; Hitchcock, 2007)), or to change the representation of specific examples in such a way that the need for them is avoided. Nevertheless, we feel that the fact that all of these features are integrated into CP-logic in a natural way makes the latter a suitable language for the study of actual causation.

In this chapter we have exploited this fact by developing a general framework for actual causation in the context of CP-logic. The essence of this framework lies in the concepts of *Intrinsicness* and *Irrelevance* on the one hand, and the extension to actual causation presented in the next chapter on the other (Definitions 3 and 15). While similar concepts could be defined in the context of structural models, their definition is much more straightforward in the temporal, non-deterministic semantics of CP-logic.

The second type of criticism claims that although the HP definition is on the right track, it requires some adjustments in order to handle certain convincing counterexamples. As a result, several authors have proposed definitions using structural equations that are variants of the HP definition (Halpern, 2015a; Hitchcock, 2001, 2007; Weslake, 2015; Woodward, 2003). Weslake (2015) offers an insightful comparison of several of these variants, concluding that none of them succeed in dealing with all counterexamples in a satisfactory manner.

The basic idea behind the HP definition and its variants formulated using structural equations is very similar to the idea behind our general definition formulated using CP-logic: construct variations of the causal model using information from the actual story, and check if there is counterfactual dependence of the effect on the candidate cause in one of these variations. Besides the use of different formal languages, the difference between both approaches is twofold. First, we use probabilities to quantify the importance of causes. As we will see in the next chapter, this proves particularly

helpful in our discussion of an extension to actual causation. The second, more fundamental, difference lies in the methodology of constructing variations of a causal model.

The HP-like definitions construct variations using so-called *structural contingencies*. A structural contingency is some set of interventions on a structural model. Different approaches differ in which structural contingencies they allow. Typically, no principled account is given of why certain structural contingencies should be allowed or not. Instead, this is decided in an *ad hoc* manner, based on whether allowing them provides the right answer for certain problematic examples. Comparisons between different approach are therefore typically also reduced to a tally of (in-)correctly handled examples.

As we have seen, in our approach on the other hand, the construction of variations is determined by the *Intrinsicness* and *Irrelevance* functions. Therefore different instantiations of our general definition can be compared directly, and can be defended by means of principled arguments in favour of particular definitions for these functions. Given the overwhelming amount of problematic examples and the conflicting intuitions that come with them, we believe a systematic approach to defining actual causation is the right way forward.

In the next chapter we turn to the second goal of our first part: to incorporate the influence of the context into judgments of actual causation.

Chapter 4

The Halpern and Hitchcock Extension to Actual Causation

Parts of this chapter were previously published as:
Beckers, S., and Vennekens, J. (2015a). Combining probabilistic, causal, and normative reasoning in cp-logic. In *12th international symposium on logical formalizations of commonsense reasoning* (pp. 32–38).

4.1 Introduction

In their forthcoming article *Graded Causation and Defaults*, Halpern and Hitchcock – HH – quite rightly observe that not only is there a vast amount of disagreement regarding actual causation in the literature, but there is also a growing number of empirical studies which show that people's intuitions are influenced to a large degree by contextual factors which up to now have been ignored when dealing with causation. For example, our judgments on two similarly modelled cases may differ depending on whether it takes place in a moral context or a purely mechanical one, or on what we take to be the default setting, or on whether we take something to be a background condition or not, etc.(Hitchcock & Knobe, 2009; Knobe & Fraser, 2008; Moore, 2009). This has led HH to develop a flexible framework that allows room for incorporating different judgments on actual causation. More specifically, in their view the difference between

cases that are modelled using similar structural models depends on which worlds we take to be more normal than others in the different contexts. Therefore their solution is to extend structural models with a normality ranking on worlds, and use it to adapt and order our judgments of actual causation in a manner suited for the particular context.

We sympathize with many of their observations, and we agree that normality considerations do influence our causal judgments. However, we find their representation of normality lacking for three reasons. First, although they emphasize the importance of distinguishing between statistical and normative normality, they use a single ranking for both. Second, they refrain from using probabilities to represent statistical normality, and instead work with a partial preorder over worlds. Third, although they stress the generality of their approach, they develop it solely for the HP-definition of actual causation (Halpern & Pearl, 2005a).

In the previous chapter we developed a general, parametrized definition of actual causation. In this chapter we use that definition to improve upon the extension to actual causation offered by HH. Using CP-logic, we are able to represent statistical normality in the usual way, i.e., by means of probabilities. As we will show, such a quantitative representation of statistical normality avoids a number of problems that HH's ordinal representation runs into. To cope with normative normality, we introduce a separate notion of norms. The result will be a more generally applicable and yet simpler approach.

The next section presents the extension to actual causation by HH. We translate their work into the CP-logic framework in Section 4.3. Section 4.4 contains a first improvement to this translation, followed by some examples and our final extension to actual causation in Section 4.5.

4.2 The original HH Extension to Actual Causation

In this section we succinctly present the graded, context-dependent approach to actual causation from (Halpern & Hitchcock, 2015). We will use the following story from (Knobe & Fraser, 2008) as our running example, as it illustrates the influence normative considerations can have on our causal attributions:

Example 5. *The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mailed them reminders that only administrators are allowed to take the pens. On Monday morning, one of the administrative assistants encounters professor Smith walking past*

the receptionist's desk. Both take pens. Later, that day, the receptionist needs to take an important message...but she has a problem. There are no pens left on her desk.

Although the problem the receptionist faces is counterfactually dependent on both the actions of the assistant and that of the professor, it turns out that people are far more inclined to judge the professor's action to be a cause than the action of the assistant.

We formally represent the relevant events on the Monday morning from our running example using CP-logic. The domain consists of the variables $Prof$ and $Assistant$, which stand for the professor respectively the assistant taking a pen, and $NoPens$, which is true when there are no pens left. The causal structure can be represented by the following CP-theory T:

$$(Prof : 0.7) \leftarrow . \tag{4.1}$$

$$(Assistant : 0.8) \leftarrow . \tag{4.2}$$

$$NoPens \leftarrow Prof \wedge Assistant. \tag{4.3}$$

The given theory summarizes all possible stories that can take place in this model. One of those is what in fact did happen that Monday morning: both the professor and the assistant take a pen, leaving the receptionist faced with no pens. The other stories consist in only the professor taking a pen, only the assistant doing so, or neither, as can be seen in the probability tree in Figure 4.1. The leftmost branch is the formal counterpart of the above story.

Like much work on actual causation, HH frame their ideas using structural equation modelling. As discussed in Section 2.2, such a model consists of a set of equations, one for each endogenous variable, that express the functional dependencies of the endogenous variables on others. As we did before, HH restrict attention to models containing two types of (Boolean) endogenous variables: the ones that deterministically depend on other endogenous variables, and those that depend directly on exogenous variables.

For example, the above CP-theory is expressed using structural equations as follows:

$$Prof := U_1 \tag{4.4}$$

$$Assistant := U_2 \tag{4.5}$$

$$NoPens := Prof \wedge Assistant. \tag{4.6}$$

HH take a *world* to be an assignment to all endogenous variables. Given a structural model, each assignment $\mathbf{u}$ to the exogenous variables determines a unique

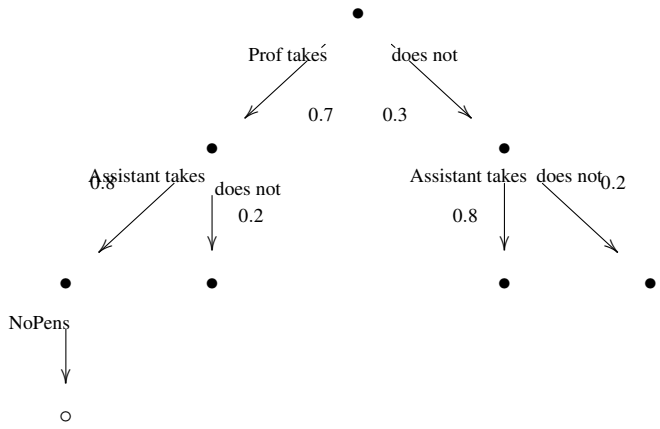


Figure 4.1: Probability tree for the Pen-vignette.

world, denoted by $s_{\mathbf{u}}$. The story from our example corresponds to the world $\{Prof, Assistant, NoPens\}$.

An extended structural model $(M, \succeq)$ consists of a structural model M together with a normality ranking $\succeq$ over worlds. This ranking is a partial pre-order informed by our – possibly subjective – judgments about what we take to be normal in this context. It is derived by considering the typicality of the values of the variables in each world. A world s is more normal than s' if there is at least one variable that takes a more typical value in s than it does in s', and no variable takes a less typical value. For the variables that depend only on other endogenous variables, things are straightforward: it is typical that these variables take the value dictated by their deterministic equation. For the other variables, i.e., those that depend only on the exogenous variables, a typicality ranking over their possible values has to be provided by the modeller.

For example, since assistants typically do take pens, the actual world $\{Prof, Assistant, NoPens\}$ is more normal than $\{Prof, \neg Assistant, \neg NoPens\}$. The latter is still more normal than $\{Prof, \neg Assistant, NoPens\}$, because here $NoPens$ violates its equation.

Typicality and normality are meant to encompass both statistical and normative judgments. Although HH make no syntactic distinction between the two kinds of normality, in the examples discussed they do differentiate between them informally. Concretely, normative judgments trump statistical judgments. For example, although professors typically do take pens, they shouldn't, and therefore $\{\neg Prof, Assistant, \neg NoPens\}$ is more normal than $\{Prof, Assistant, NoPens\}$. By contrast, we will make a formal distinction between the two forms of normality,

because in this manner we can incorporate information regarding both.

Say we have a story, i.e., an assignment to all variables, such that C and E happen in it. In order to establish whether C is a cause of E, any definition in the counterfactual tradition restricts itself to some particular set of counterfactual worlds in which $\neg C$ holds and checks whether also $\neg E$ holds in these worlds. If this set contains a world which serves to justify that C is indeed a cause of E (i.e., one in which E is false), then HH call such a world a *witness* of this. HH adapt a given definition of actual causation using the normality ranking to disallow worlds that are less normal than the actual world, in order to reflect the influence of normality on possible causes.

Definition 7. *[HH-extension of actual causation.] Given are an extended structural model $(M, \succeq)$ and exogenous assignment $\mathbf{u}$, such that both C and E hold in $s_{\mathbf{u}}$. C is an HH-actual cause of E in $(M, \succeq, \mathbf{u})$ if C is an actual cause of E in $(M, \mathbf{u})$ when we consider only witnesses w such that $w \succeq s_{\mathbf{u}}$.*

Since we have a ranking on the normality of worlds, this definition straightforwardly leads to an ordering between different causes indicating the strength of the causal relationship by looking at the highest ranked witness for a cause, which is called its *best witness*.

As mentioned, in case of our example, we have

$$\{\neg Prof, Assistant, \neg NoPens\} \succ \{Prof, Assistant, NoPens\} \succ \{Prof, \neg Assistant, \neg NoPens\}$$

Since the actual world is in the middle, the first world can serve as a witness for $Prof$ being a cause, but the last world may not be used to judge $Assistant$ to be a cause. Hence using a normality ranking HH are able to distinguish between the actions of the professor and that of the assistant, in line with the observations made by Knobe and Fraser (2008).

4.3 The HH Extension in CP-logic

We proceed with translating the HH-extension of actual causation into CP-logic. To get there, we will translate one by one all of the required concepts.

HH use the HP-definition (Halpern & Pearl, 2005a) (presented in Section 2.2.1) as a working definition to illustrate their extension to actual causation. However, they stress the generality of their approach, and mention that one could for example apply it to Hall's definition (2007), which we reformulated as an instantation of our general definition in Section 3.7. To keep things simple, we will use Dependence (Definition 2) as our definition of actual causation throughout the examples, and thus take T^* to

simply be T^b. Note however that our approach can be applied to any instantiation of our general definition.

As explained in Section 2.4, a structural model M in the HH-setting corresponds to a CP-theory T: a direct dependency on the exogenous variables results in a non-deterministic, vacuous law (such as (4.1) and (4.2)), while a dependency on only endogenous variables results in a deterministic law (such as (4.3)).

A world s described by M then corresponds to a branch b – or to be more precise, the leaf of a branch – of a probability tree of T. In Definition 7, HH restrict attention to those worlds that are at least as normal as the actual world. The CP-logic equivalent of this will be the *normal refinement* of T according to b, which is a theory that describes those stories which are at least as normal as b.

First we introduce two operations on CP-laws, corresponding to the two different interpretations of normality. Assume at some point a CP-law r was applied in b, choosing the disjunct r_b that occurs in its head.

On a probabilistic reading, an alternative application of r is at least as normal as the actual one if a disjunct is chosen which is at least as likely as r_b. Therefore the *probabilistically normalized* refinement of r according to b – denoted by $r^{PN(b)}$ – consists in r without all disjuncts that have a strictly smaller probability than r_b. Remain the laws that were not applied in b. In line with the understanding of normality from HH, we choose to handle these such that they cannot have effects which result in a world less normal than the actual world. Thus we remove those disjuncts that are false in the leaf of b and have a probability lower than 0.5. Further, say the total probability of the removed disjuncts in some law is p, then we renormalize the remaining probabilities by dividing by $1-p$. (Unless $p=1$, then we simply remove the law.)

Definition 8. *Given a theory T, and a story b, we define the* probabilistically normalized refinement *of T according to b as $T^{PN(b)} := \{r^{PN(b)} | r \in T\}$.*

In case of our example, $T^{PN(b)} = T^b$, shown earlier.

A second reading of normality considers not what did or could happen, but what ought to happen. To allow such considerations, we extend CP-logic with *norms*. Everything that can possibly happen is described by the CP-laws of a theory, hence we choose to introduce *prescriptive* norms in addition to *descriptive* CP-laws. These take the form of alternative probabilities for the disjuncts in the head of a law, which represent how the law should behave. (A more general approach could be imagined, but for the present purpose this extension will suffice.) These probabilities will be enclosed in curly braces, and have no influence on the actual behaviour of a theory. However, if we wish to look at how the world should behave, we can enforce the norms by replacing the original probabilities of a theory T with the normative ones. For example, extending law (4.1) with the norm that the professor shouldn't take pens, even though he often does, gives

$(Prof : 0.7\{0\}) \leftarrow$. The *normatively normalized* refinement – denoted by $r^{NN(b)}$ – is then given by the CP-law $(Prof : 0) \leftarrow$. To properly capture the HH definition, our normalized theory should allow all worlds that are at least as normal as b, including of course b itself. For this reason, we here restrict attention to norms with a probability p so that $0 < p < 1$, because a norm $p = 0$ or $p = 1$ could make the actual world b impossible. This restriction is lifted in our own proposal in Section 4.5.1. In the meanwhile we shall use $(Prof : 0.7\{p\}) \leftarrow$, where p is some small probability, e.g., 0.01.

Definition 9. *Given an extended theory T, i.e., a theory also containing norms, we define the* normatively normalized refinement *of T as* $T^{NN} := \{r^{NN} | r \in T\}$.

We can combine both senses of normality, as follows:

Definition 10. *Given an extended theory T, and a story b, we define the* normal refinement *of T according to b as* $T^{Normal(b)} := (T^{NN})^{PN(b)}$.

The fact that we first take the normative normalization, before taking the probabilistic one, corresponds to the implicit assumption made by HH that normative judgments trump probabilistic ones. To see the difference, imagine a story in which the professor did not take a pen. As we have now defined it, $(Prof : 0.7\{p\}) \leftarrow$ then normalises to $\leftarrow$, i.e., $Prof$ becomes impossible. If we were to reverse the order of normalization, it would normalise to $(Prof : p) \leftarrow$.

The normal refinement according to b is constructed out of T by eliminating all the disjuncts with values of variables that are less normal than the values from b, and thus it allows precisely those stories which are at least as normal as b, in the sense of Definitions 8 and 9. In case of our example, $T^{Normal(b)}$ is given by:

$$(Prof : p) \leftarrow . \tag{4.7}$$

$$Assistant \leftarrow . \tag{4.8}$$

$$NoPens \leftarrow Prof \wedge Assistant. \tag{4.9}$$

To be able to prove that this is indeed the CP-logic equivalent of Definition 7, we still need to explain how we get from a given extended structural model to an extended CP-theory. The normality ranking is derived by considering what is typical for those variables depending directly on the exogenous variables, i.e., those that we represent by $X : p \leftarrow$. HH use statements that take the form: "it is typical for the variable X to be **true**", or "it is typical for it to be **false**". In CP-logic this becomes: $p > 0.5$, and $p < 0.5$ respectively. A statement of the form: "it is more typical for X to be **true** than for Y to be **true**" translates in an ordering on the respective probabilities. If the typicality statement is of the normative kind, then it is best represented by norms in

CP-logic. Thus if there is a norm regarding X then the law will take the extended form $X : p\{q\} \leftarrow$.

In Definition 7, we have that a world is an acceptable witness only if it belongs to the set of worlds allowed by the definition of actual causation that is being used, and it is at least as normal as the actual world. Similarly, we need to limit the stories allowed by Definition 3 – which are described by $T^*|do(\neg C)$ – to those stories which are at least as normal as b. We would like to do this in exactly the same manner as we did for T, i.e., by looking at $(T^*|do(\neg C))^{Normal(b)}$. However, by applying the intervention $do(\neg C)$, which removes C from the head of any law in T in which it appears, we lose all information on the (ab)normality of $\neg C$. As this information should be taken into account as well, we have to incorporate it somehow.

One solution to do so is by simply factoring in $\mathbf{P}_{T^{Normal(b)}}(\neg C)$. In Section 4.4 we present an argument in favour of this solution. Since HH do not quantify normality, this solution is not available to them. Instead, they use the normal refinement of those laws instead. Concretely, denote by $R(C)$ those laws from T which have C in their head. We denote by $T^{**}(C)$ the theory that is identical to $T^*|do(\neg C)$ regarding all laws that are not in $R(C)$, and containing the normal refinements of all laws in $R(C)$. This complication disappears further on, when we consider the first solution mentioned.

Definition 11. *Given an extended theory T, a story b such that C and E hold in its leaf, and the theory $T^* = [T \setminus (Irr(Con) \cup Int(Con))] \cup Int(Con)^b$, as described in Definition 3. We define the* normal refinement *of T^* according to b and C as $T^{Normal(b)*}(C) := (T^{**}(C))^{Normal(b)}$.*

This leads us to the following formulation of the HH-approach in CP-logic.

Definition 12. *[HH-CP-logic-extension of actual causation] Given an extended theory T, and a branch b such that both C and E hold in its leaf. We define that C is an* HH-CP-logic-actual cause *of E in (T,b) if $\mathbf{P}_{T^{Normal(b)*}(C)}(\neg E \wedge \neg C) > 0$.*

For example, if we are considering whether *Assistant* caused *NoPens* in our story, the theory $T^{Normal(b)*}(Assistant)$ is given by T^b. (Recall that we take Dependence as our working definition.)

$$Prof \leftarrow . \tag{4.10}$$

$$Assistant \leftarrow . \tag{4.11}$$

$$NoPens \leftarrow Prof \wedge Assistant. \tag{4.12}$$

This gives $P(\neg NoPens \wedge \neg Assistant) = 0$.

On the other hand, the theory $T^{Normal(b)*}(Prof)$ is given by:

$$(Prof : p) \leftarrow . \tag{4.13}$$

$$Assistant \leftarrow . \tag{4.14}$$

$$NoPens \leftarrow Prof \wedge Assistant. \tag{4.15}$$

This gives $P(\neg NoPens \wedge \neg Prof) = 1 - p$.

Thus $Prof$ is judged to be a strong cause of *NoPens*, whereas *Assistant* isn't a cause at all, in line with the empirical results from (Knobe & Fraser, 2008). Note that it is only by using the normative probabilities rather than the statistical ones that we get the correct response for $Prof$.

In Section 3.2.1 we gave a general definition of actual causation in terms of CP-logic. In order to finish the translation from extended structural models to extended CP-logic, we assume that the definition of actual causation being used is such that just as the three definitions from Hall discussed in Section 3.7 and 3.4, it can be translated from structural models into the framework of our general definition. We now show that Definition 12 is indeed the correct translation of the HH approach from structural models to CP-logic.

Theorem 7. *C is an HH-actual cause of E in an extended model and exogenous assignment $(M, \succeq, \mathbf{u})$ iff C is an HH-CP-logic-actual cause of E in (T, b), where (T, b) is derived from $(M, \succeq, \mathbf{u})$ in the sense described above.*

To facilitate the proof of Theorem 7, we introduce the following lemma.

Lemma 2. *Given an extended model and exogenous assignment $(M, \succeq, \mathbf{u})$, and a theory and branch (T, b) that are derived from $(M, \succeq, \mathbf{u})$ in the sense described in Section 4.3. Then for any world w, and a branch d of a probability tree from T that corresponds to it, it holds that $w \succeq s_{\mathbf{u}}$ iff d occurs in a probability tree of $T^{Normal(b)}$.*

Proof. We know that b is a branch in a probability tree from T such that $Leaf_b$ has the same assignment as $s_{\mathbf{u}}$. Recall that T consists of two categories of laws. First there are those corresponding to the equations for the endogenous variables which depend on other endogenous variables, which are deterministic and thus re-appear in $T^{Normal(b)}$ unchanged. Second there are those corresponding to the endogenous variables which directly depend on the exogenous variables, which take the form $X : p\{q\} \leftarrow$, where the second probability need not be present.

Assume we have a world w such that $w \succeq s_{\mathbf{u}}$. Any world that satisfies the equations of M follows deterministically from an assignment to all exogenous variables. As

$s_{\mathbf{u}}$ is a world that satisfies the equations, and w is at least as normal, it also satisfies the equations. Hence there is an exogenous assignment $\mathbf{u}'$ which determines w. In CP-logic, such an assignment corresponds to choosing particular disjuncts in the heads of all laws from the second category.

Concretely, this means that for each law/equation of the second category, the value of the corresponding variable X is at least as typical in $w = s_{\mathbf{u}'}$ as it is in $s_{\mathbf{u}}$. Denote by X_w and X_s the values X takes in the worlds w and $s_{\mathbf{u}}$ respectively. By construction of $T^{Normal(b)}$, the disjuncts which are at least as typical as X_s – in the normative sense where applicable, in the statistical sense elsewhere – still appear in the law for X in $T^{Normal(b)}$, and hence can be chosen when this law is applied. Therefore the branches corresponding to w from the probability trees of T also appear in the probability trees of $T^{Normal(b)}$, be it that the values of the probabilities may have changed.

Now assume we have a branch d corresponding to a world w, that occurs in a probability tree of $T^{Normal(b)}$. We can simply reverse the correspondence between the choices of disjuncts and an exogenous assignment, to obtain that $w \succeq s_{\mathbf{u}}$.

□

Now we prove Theorem 7.

Proof. We begin with the implication from left to right. So assume we have an extended model and exogenous assignment $(M, \succeq, \mathbf{u})$, such that C and E hold in $s_{\mathbf{u}}$, and there is at least one witness w of C being an actual cause of E in $(M, \mathbf{u})$ such that $w \succeq s_{\mathbf{u}}$.

Recall that we assume the definition of actual causation at hand can be translated from structural models into an instantiation of our general definition. So we get that C is an actual cause of E in (T, b), and more specifically that any branch d that corresponds to w is a witness of this. Thus d appears in a probability tree of $T^*|do(\neg C)$.

By Lemma 2, we know that such a branch d also appears in a probability tree of $T^{Normal(b)}$.

We look separately at the two options regarding $R(C)$. First we assume that C is determined directly by the exogenous variables, meaning that $R(C)$ consists of a single non-deterministic law, say $r(C)$. Since d occurs in a tree of $T^{Normal(b)}$, and $\neg C$ holds in it, the empty disjunct remains present in the normal refinement of $r(C)$. By definition, $T^{**}(C)$ is simply $T^*|do(\neg C)$ with the normal refinement of $r(C)$. Therefore d also occurs in a tree of $T^{**}(C)$.

Second, assume the laws in $R(C)$ are deterministic. Since d occurs in a tree of $T^{Normal(b)}$, which obviously contains C in the head of any law $r(C) \in R(C)$, the body for $r(C)$ cannot be satisfied in d. Thus all laws $R(C)$ are irrelevant to d. Since $T^{**}(C)$

and $T^*|do(\neg C)$ are identical but for $R(C)$, we can again conclude that d also occurs in a tree of $T^{**}(C)$.

So in all cases we have that d occurs both in a tree of $T^{Normal(b)}$, and in a tree of $T^{**}(C)$. This implies that the disjuncts chosen in the laws applied in d occur in the versions these laws take in both of these theories, with possibly different but strictly positive probabilities. Note that every law from $T^{Normal(b)*}(C)$ either takes the form it has in $T^{**}(C)$ or it takes the form it has in $T^{Normal(b)}$. Therefore d also appears in $T^{Normal(b)*}(C)$. It being a witness, $\neg C$ and $\neg E$ hold in it, and thus the stated probability is strictly positive.

Now we continue with the reverse implication. Assume we have an extended theory T, a story b such that C and E hold in it, and $\mathbf{P}_{T^{Normal(b)*}(C)}(\neg E \wedge \neg C) > 0$. This implies the existence of a branch d in $T^{Normal(b)*}(C)$ such that both $\neg C$ and $\neg E$ holds.

Say r is a law from $T^{Normal(b)*}(C)$. If r is intrinsic and $r \notin R(C)$, it is deterministic, containing the single (possibly empty) disjunct r_d with associated probability 1. As r_d was the actual choice from b, by construction r_d also appears in the normal refinement of r, although the probability may be different. However, as long as we do not have strict norms, i.e., norms where p or q is 1, this probability will be strictly positive. A strict norm means that a violation of it is considered entirely abnormal, analogous to the occurrence of an event with zero probability. Since HH treat norms identical to statistical normality, and since the actual world was possible, it follows that the actual world is not entirely abnormal. Hence even if r_d was a violation of a norm, it will not have been a strict norm. (Our final definition from Section 4.5.1 does allow for strict norms.) Thus, we conclude that r_d occurs in the head of the versions of the law r we find in both T^* and $T^{Normal(b)}$. Because $r \notin R(C)$, we can say the same about $T^*|do(\neg C)$.

If r is not intrinsic and $r \notin R(C)$, it contains all of its original disjuncts when it occurs in T^*. Therefore it takes the same form in $T^{Normal(b)*}(C)$ as it does in $T^{Normal(b)}$. Again we conclude that r_d occurs in the head of the versions of the law r we find in each of T^*, $T^{Normal(b)}$ and $T^*|do(\neg C)$.

This leaves us to consider the laws in $R(C)$. By definition, $T^{Normal(b)*}(C)$ contains the same version of these laws as $T^{Normal(b)}$. From this and the previous paragraphs we can already conclude that any branch occuring in a tree of $T^{Normal(b)*}(C)$ also occurs in a tree of $T^{Normal(b)}$. More specifically this holds for d. Thus by Lemma 2, it holds for the corresponding world w that $w \succeq s_u$.

If the body for each $r \in R(C)$ is false in d, then the precise form of the head of r is irrelevant for d. As the head of each $r \in R(C)$ is the only difference between $T^{**}(C)$ and $T^*|do(\neg C)$, we can again conclude that d also occurs in $T^*|do(\neg C)$.

Leaves us to consider the case that there is some $r \in R(C)$ for which the body is true in

d. From the fact that d – in which $\neg C$ holds – occurs in $T^{Normal(b)}$, we can infer that r is a non-deterministic law, and thus the only member of $R(C)$. Taken together with the knowledge that the disjunct containing C was chosen in b, it follows that the normal refinement of r contains both C and the empty disjunct in its head. Furthermore, in d the empty disjunct was chosen. These observations taken together imply that the disjunct of r chosen in d occurs in the head of the versions of r we find in both $T^{**}(C)$ and $T^*|do(\neg C)$. Once more we conclude that d also occurs in $T^*|do(\neg C)$.

Thus d is a witness for C being an actual cause of E in (T,b). Therefore the world w corresponding to d is a witness for C being an actual cause of E in (M,u). Together with the fact that $w \succeq s_u$, the conclusion follows.

□

4.4 The Importance of Counterfactuals

We mentioned earlier that one criterion for a story to be normal was that it respects the laws/equations. On the other hand definitions of actual causation look at counterfactual stories resulting from an intervention, namely $do(\neg C)$, which violates the laws. Following HH, Definition 12 tries to circumvent the use of this intervention by only allowing witnesses in which $\neg C$ happens to hold. However, it may be the case that this condition actually eliminates all potential witnesses. When this happens, counterintuitive results may follow. We illustrate what goes wrong by using the following theory:

$$(A : 0.1) \leftarrow . \qquad E \leftarrow C.$$

$$C \leftarrow A. \qquad E \leftarrow \neg A.$$

Consider the story where first A occurs, followed by C and E. Intuitively, C is a strong cause of E, because it is an atypical phenomenon ($P(C) = 0.1$) without which E would not have occurred. The law with A in its head is intrinsic, and thus $T^{Normal(b)*}(C)$ is:

$$A \leftarrow . \qquad E \leftarrow C.$$

$$C \leftarrow A. \qquad E \leftarrow \neg A.$$

Applying the definition, we get that $\mathbf{P}_{T^{Normal(b)*}(C)}(\neg E \wedge \neg C) = 0$, giving the absurd result that C is not a cause of E at all. The problem lies in the fact that in its current form we only allow stories containing $\neg C$ in the usual, lawful way, rather than stories

which contain $\neg C$ as a result of the intervention $do(\neg C)$. The problem remains if we use the HP-definition – as HH does – instead of our working definition.

We pointed out above that there is another solution to take into account the normality of $\neg C$. As a first step, we look instead at $\mathbf{P}_{T^{Normal(b)*}(C)}(\neg E|do(\neg C))$, so that we re-establish the counterfactual nature of our definition. (As $T^{**}(C)|do(\neg C) = T^*|do(\neg C)$, this is equivalent to $\mathbf{P}_{(T^*)^{Normal(b)}}(\neg E|do(\neg C))$, which no longer mentions the artificial theory $T^{**}(C)$.) However, by making this move we no longer take into account the (ab)normality of C itself, whereas research shows extensively that causal judgments regarding an event are often influenced by how normal it was (Hitchcock & Knobe, 2009; Kahneman, 1986; Knobe & Fraser, 2008). (This effect is not limited to normative contexts. For example, the lighting of a match is usually judged a cause of a fire, whereas the presence of oxygen is considered so normal that it isn't.) Hence as a second step we factor in this normality, which is expressed by $\mathbf{P}_{T^{Normal(b)}}(\neg C)$.

Definition 13. *[First refinement of Definition 12] Given an extended theory T, and a branch b such that both C and E hold in its leaf. We define that C is an actual cause of E in (T,b) if* $\mathbf{P}_{(T^*)^{Normal(b)}}(\neg E|do(\neg C)) * \mathbf{P}_{T^{Normal(b)}}(\neg C) > 0$.

As the following theorem shows, our new choice only makes a difference in a limited set of cases.

Theorem 8. *If $R(C)$ contains a non-deterministic law or $\mathbf{P}_{T^{Normal(b)}}(\neg C) = 0$, then*

$$\mathbf{P}_{T^{Normal(b)*}(C)}(\neg E \wedge \neg C) = \mathbf{P}_{(T^*)^{Normal(b)}}(\neg E|do(\neg C)) * \mathbf{P}_{T^{Normal(b)}}(\neg C)$$

Proof. First we examine the case where $\mathbf{P}_{T^{Normal(b)}}(\neg C) = 0$. This implies that the right-hand side of the equation is 0. Also, any branch from a tree $T^{Normal(b)*}$ occurs as well in a tree of $T^{Normal(b)}$, so $\mathbf{P}_{T^{Normal(b)*}}(\neg C) = 0$ and the left-hand side is also equal to 0.

This leaves us to consider the case where $\mathbf{P}_{T^{Normal(b)}}(\neg C) > 0$ and the unique $r(C) \in R(C)$ is non-deterministic.

In this case $\mathbf{P}_{T^{Normal(b)*}}(\neg C) = \mathbf{P}_{T^{Normal(b)}}(\neg C)$, so we have:

$\mathbf{P}_{T^{Normal(b)*}}(\neg E \wedge \neg C) = \mathbf{P}_{T^{Normal(b)*}}(\neg E \wedge \neg C) * \mathbf{P}_{T^{Normal(b)}}(\neg C)/\mathbf{P}_{T^{Normal(b)*}}(\neg C) = \mathbf{P}_{T^{Normal(b)*}}(\neg E|\neg C) * \mathbf{P}_{T^{Normal(b)}}(\neg C)$

Further, conditioning on $\neg C$ when C only occurs in a vacuous non-deterministic law, is identical to looking at the intervention $do(\neg C)$, thus the list of equalities continues:

$= \mathbf{P}_{T^{Normal(b)*}}(\neg E|do(\neg C)) * \mathbf{P}_{T^{Normal(b)}}(\neg C)$. Also, $T^{Normal(b)*}|do(\neg C) = (T^*)^{Normal(b)*}|do(\neg C)$, which brings us to the desired conclusion.

□

If there is a deterministic $r \in R(C)$ and $\mathbf{P}_{T^{Normal(b)}}(\neg C) > 0$, as in the example shown, then contrary to the left-hand side of the equation, the proposed adjustment on the right-hand side of the equation gives the desired result $1 * 0.9 = 0.9$.

4.5 The Importance of Probabilities

Because the HH-approach lacks the quantification of normality offered by probabilities, they dismiss entirely all witnesses that are less normal than the actual world. A direct consequence is that any typical event – i.e., $P > 0.5$ – is never a cause, which is quite radical. By using probabilities, this qualitative criterion is no longer necessary: less normal witnesses simply influence our causal judgment less. Further, HH order causes solely by looking at the best witnesses. We now present an example which illustrates the benefit of both abandoning their criterion, and aggregating the normality of witnesses to order causes, without sacrificing the influence of normality.

> Imagine you enter a contest. If a 10-sided die lands 1, you win a car. If not, you get a 100 more throws. If all of them land higher than 1, then you also win the car. The first throw lands 1, and you win the car.

It's hard to imagine anyone objecting to the judgment that the first throw is a cause of you winning the car. Yet that is exactly what we get when applying either Definition 12 or the improved Definition 13. The following theory T describes the set-up of the contest, where $Throw(i, j)$ means that the i-th throw landed j or smaller.

$$(Throw(1,1) : 0.1) \leftarrow .$$

$$(Throw(2,1) : 0.1) \leftarrow \neg Throw(1,1).$$

$$(Throw(3,1) : 0.1) \leftarrow \neg Throw(1,1) \wedge \neg Throw(2,1).$$

$$\ldots$$

$$WinCar \leftarrow Throw(1,1).$$

$$WinCar \leftarrow \neg Throw(2,1) \wedge \ldots \wedge \neg Throw(100,1).$$

The normal refinement of T according to the story is given by:

$$(Throw(1,1) : 0.1) \leftarrow .$$

$$WinCar \leftarrow Throw(1,1).$$

$$WinCar \leftarrow \neg Throw(2,1) \wedge \ldots \wedge \neg Throw(100,1).$$

We get that $\mathbf{P}_{(T^*)^{Normal(b)}}(\neg WinCar|do(\neg Throws(1,1))) = 0$, and thus $Throws(1,1)$ is not a cause of $WinCar$. In terms of HH: although $\neg Throw(1,1) \wedge \neg Throw(2,1) \wedge ... \wedge \neg Throw(100,1) \wedge WinCar$, is very unlikely, it is the only candidate witness. To see why, recall that a witness needs to have $\neg Throws(1,1)$, and should be at least as normal as the actual world. In every other world with $\neg Throws(1,1)$, at least one of the $Throws(i,1)$ is true, and hence it is less normal. But in a witness it should hold that $\neg WinCar$, so there is no witness for $Throws(1,1)$ being a cause of $WinCar$.

We can fix this problem by considering the theory $(T^*)^{NN}$ instead of $(T^*)^{Normal(b)}$. This leads us to another refinement of our original definition:

Definition 14. *[Second refinement of Definition 12] Given an extended theory T, and a branch b such that both C and E hold in its leaf. We define that C is an actual cause of E in (T,b) if $\mathbf{P}_{(T^*)^{NN}}(\neg E|do(\neg C)) * \mathbf{P}_{T^{Normal(b)}}(\neg C) > 0$.*

The theory $(T^*)^{NN}$ in this case is simply equal to T, but for the first law being $Throw(1,1) \leftarrow$. Hence the probability of not winning the car given that the first throw does not land 1 is pretty much 1, and the value in the equation becomes approximately 0.9, indicating $Throw(1,1)$ to be a very strong cause of $WinCar$.

Note that we only obtain this high value because the probability $\mathbf{P}_{(T^*)^{NN}}$ aggregates the probabilities of all witnesses. If we would instead follow HH in considering only the best witness (in this case the story with $\neg Throw(1,1) \wedge Throw(2,1) \wedge \neg Throw(3,1) \wedge ... \wedge \neg Throw(100,1)$), we would obtain the much lower and less intuitive probability of 0.09.

Now imagine the same story, with a slight variation to the rules of the contest: you win the car on the first throw if the die lands anything under 7. Hence the first head changes to $Throw(1,6) : 0.6$, making it a typical outcome. Therefore the first law becomes deterministic in $T^{PN(b)}$, giving that $\mathbf{P}_{T^{Normal(b)}}(\neg Throw(1,6)) = 0$, which again results in the counterintuitive judgment that the first throw in no way caused you to win the car.

We therefore suggest to use T^{NN} in the second factor of the inequality rather than $T^{Normal(b)}$, making use of the gradual measurement offered by probabilities. Applying this idea to the example, we get the result that $Throw(1,6)$ has causal strength 0.4. This value is smaller than before, because the cause is now less atypical.

4.5.1 The final definition

This brings us to our final extension to a definition of actual causation.

Definition 15 (Extension of actual causation)**.** *Given an extended theory T, and a branch b such that both C and E hold in its leaf. We define that C is an actual cause of E in (T,b) if and only if* $\mathbf{P}_{(T^*)^{NN}}(\neg E|do(\neg C)) * \mathbf{P}_{T^{NN}}(\neg C) > 0$.

4.6 Conclusion

In this chapter we showed how our general definition lends itself to an extension of actual causation that continues upon the work of Halpern and Hitchcock (2015). It incorporates the main points raised by Halpern and Hitchcock: (1) it allows normative considerations and (2) is able to factor in the normality of the cause given the context. This extension is useful in normative disciplines, such as law and ethics, and takes into account the context sensitivity of causal judgements suggested by recent findings in experimental psychology (Hitchcock & Knobe, 2009; Knobe & Fraser, 2008; Moore, 2009). Our account improves on the HH account in several ways:

- By using CP-logic our account can also be applied to non-deterministic examples.
- Separating normative from statistical normality allows for a more accurate description of the domain.
- Since we no longer refer to the actual world in the second factor, we can use strict norms.
- The reader may verify that our approach is able to deal with all of the examples given by Halpern and Hitchcock (2015) equally well as Definition 12.
- It can also properly handle the examples from Sections 4.4 and 4.5, as opposed to Definition 12.

Chapter 5

Problems with BV12 and Hall07

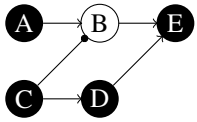
5.1 Introduction

In this chapter we conclude the first part of this work by presenting several problems facing the Hall07 and BV12 definitions. First we focus on a flaw in the Hall07 definition that makes it sensitive to irrelevant details of a story. Second, we present two problems for the BV12 definition, the first of which affects the Hall07 definition as well.

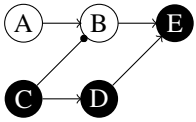
5.2 Problems with Hall07

5.2.1 *Early Preemption*

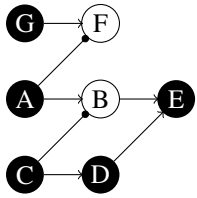
Anyone familiar with the literature on actual causation knows that cases of *Early Preemptio*n figure prominently in it, hence it forms the first challenge for any definition to overcome. We present three examples, only the first of which is a case of *Early Preemption* proper, and the other two are variations.



EP



Variant 1



Variant 2

We already saw the first two diagrams in Chapter 2, to illustrate the workings of neuron diagrams. There we pointed out that the first is a case of *Early Preemption*, because the chain from A to E is preempted from running to completion. In the second diagram, A is in its default state, and hence there is nothing to preempt. Variant 2 is identical to EP, except that we added a variable F which depends on A and some new variable G. Intuitively, this change does not affect the causal relation between C and E in any way. As we will see, Hall07 (and thus also Hall's original definition) violate this intuition.

Theory for EP and Variant 1

$(A : p) \leftarrow .$

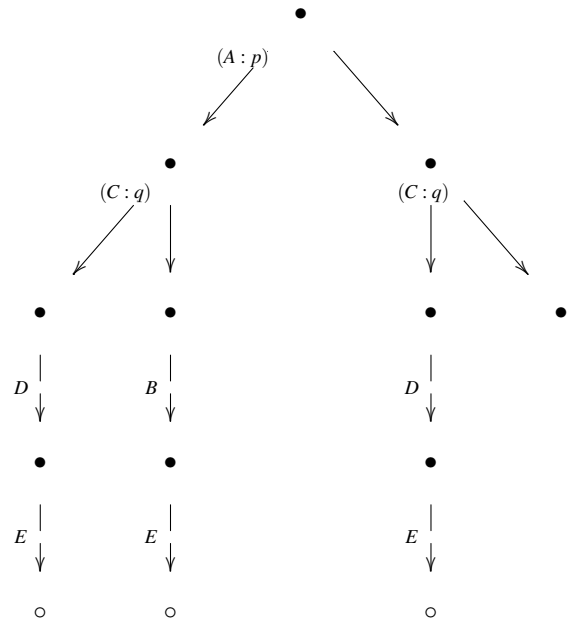
$(C : q) \leftarrow .$

$B \leftarrow A \wedge \neg C.$

$D \leftarrow C.$

$E \leftarrow B.$

$E \leftarrow D.$



On the left we show the CP-theory for EP and Variant 1, and on the right is one of its probability trees. EP corresponds to the leftmost branch, whereas Variant 1 corresponds to the third branch from the left. We apply the BV12 and Hall07 definitions to these

examples, Variant 2 can be treated in a similar manner. First we modify the original theories to take into account the actual story.

Hall07 EP $T^*|do(\neg C)$

$(A : p) \leftarrow .$

$B \leftarrow A \wedge \neg C.$

$D \leftarrow C.$

$E \leftarrow B.$

$E \leftarrow D.$

BV12 EP $T^*|do(\neg C)$

$A \leftarrow .$

$B \leftarrow A \wedge \neg C.$

$D \leftarrow C.$

$E \leftarrow B.$

$E \leftarrow D.$

Hall07 and BV12 Variant 1 $T^*|do(\neg C)$

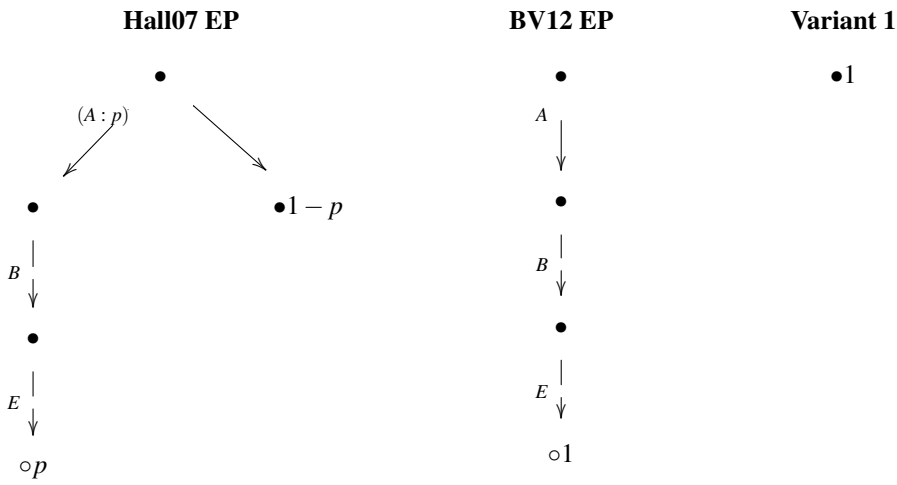
$B \leftarrow A \wedge \neg C.$

$D \leftarrow C.$

$E \leftarrow B.$

$E \leftarrow D.$

To these theories correspond the following probability trees, where we have added the probabilities for ending up in each leaf node:



Our general definition of actual causation tells us that C caused E iff $\mathbf{P}_{T^*}(\neg E|do(\neg C)) > 0$. The values of the respective probabilities are shown in Table 5.1.

Both definitions judge C to be a full cause of E in Variant 1, as both of them hold fixed the fact that A takes on its default value. Likewise, both definitions hold A fixed at its actual value in Variant 2 and therefore judge C not to be a cause of E. The BV12 definition applies the same reasoning to *Early Preemption*, namely it fixes the variable A to its actual value, and checks for counterfactual dependency of E on C.

Table 5.1: $\mathbf{P}_{T^*}(\neg E|do(\neg C))$

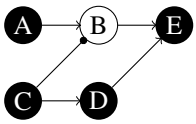
Story	Hall07	BV12
EP	$1-p$	0
Variant 1	1	1
Variant 2	0	0

The Hall07 definition on the other hand is more tolerant for *Early Preemption* than it is for Variant 2, ignoring A's actual value when considering counterfactual scenario's for the former. Yet Variant 2 represents exactly the same causal structure and story as in *Early Preemption*, except for the addition of the extra variables G and F, the latter of which depends on A not firing. Intuitively these additional variables are of no importance regarding the causal relation between C and E, so one should judge it identical in both examples, contrary to Hall07. It speaks in BV12's favour that it does not make this mistake.

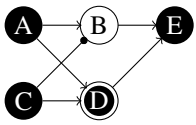
Cases of *Early Preemption* are typically considered to exhibit causation, contrary to the verdict of the BV12 definition. We will defend our verdict at length in Section 6.6, by invoking the distinction between deterministic and non-deterministic counterfactual dependence. (Concretely, the distinction between examples where the backup mechanism from B to E is expressed using a deterministic or a non-deterministic law.)

5.2.2 *Switch*

The fact that the Hall07 definition is very sensitive to the addition of extra variables and dependencies, even when intuitively these variables do not influence the actual story in any way, is not limited to the case of *Early Preemption*. Hitchcock (2009) present six counterexamples to Hall's definition, most of which are based on this sensitivity. Before providing an illustration we first introduce an example of a so-called *Switch*, which is similar to *Early Preemption*, and yet also quite different. (The double circle in D means it fires iff both C and A fire.)



EP



Switch

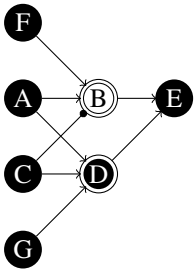
Cases of switching causation are described in (Hall, 2000, 2007; Halpern & Pearl, 2005a), amongst others. The idea is that a common cause A can activate two causal chains, or mechanisms, which both lead to an outcome E, and that the switch C determines the unique mechanism which in fact does so. The fact that D also depends on A, is the only difference between *Early Preemption* and *Switch*, which translates into the following almost identical CP-theories:

Theory for EP	**Theory for Switch**
$(A : p) \leftarrow .$	$(A : p) \leftarrow .$
$(C : q) \leftarrow .$	$(C : q) \leftarrow .$
$B \leftarrow A \wedge \neg C.$	$B \leftarrow A \wedge \neg C.$
$D \leftarrow C.$	$D \leftarrow C \wedge A.$
$E \leftarrow B.$	$E \leftarrow B.$
$E \leftarrow D.$	$E \leftarrow D.$

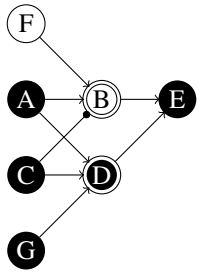
We leave it to the reader to verify that the BV12 definition treats both *Early Preemption* and *Switch* identically, judging that C is not a cause of E in either. The Hall07 definiton on the other hand does distinguish between both, only calling C a cause in the first. The motivation for Hall to get this answer is that usually in the literature informal examples labelled *Early Preemption* are such that intuitively C is a cause of E, whereas in examples labelled *Switch* intuitively C does not cause E.

We will explore the relation between these two examples in detail in Part II. We present them here merely to explain why the Hall07 definition gets into trouble: it wants to distinguish between the very similar diagrams for *Early Preemption* and *Switch*, and therefore has to focus on distinctions which turn out to be inessential. Our Variant 2 of *Early Preemption* was a first illustration, Hitchcock (2009) offers several others. One of those is a variation on *Switch*.

5.2.3 Variant of *Switch*



Variant of *Switch*: Actual situation



Variant of *Switch*: Reduction

The variation considered by Hitchcock is the result of taking the diagram for *Switch*, and adding some further dependencies, indicated by F and G. We could imagine these to represent some details which were left implicit in the original diagram. Intuitively, as with our variant of *Early Preemption*, "This extra detail does nothing to change our causal judgment" (Hitchcock, 2009, p. 395). Yet if we apply Hall's definition, we see that contrary to *Switch*, here there is a reduction of the actual situation in which there is a counterfactual dependence of E on C, making C a cause of E. Since the causal structure is *simple* in the sense defined in Chapter 3, the same holds for the Hall07 definition.

5.3 Problems with BV12

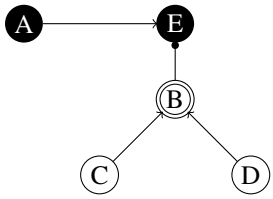
We now present two problematic examples for the BV12 definition.

5.3.1 *Symmetric Overdetermination by Omission*

Besides the example from the previous section, Hitchcock (2009) also gives some alleged counterexamples to Hall's definition involving *causation by omission*, i.e., examples where a variable remaining in its default state is a candidate cause. In these examples, intuitively $\neg C$ is a cause of E, yet applying Hall's definition does not give this result.

Looking back at Hall's definition from Section 3.7, one sees that his definition does not allow for causation by omission: the candidate cause has to be an *event*, meaning a variable taking on its deviant value. Therefore strictly speaking Hitchcock's argument is not valid, as Hall's orginal definition cannot be applied here. However Hitchcock

also presents variations of these examples where the candidate cause is an event, hence this is of minor importance. In fact, the example can be taken to offer an argument in favour of accepting causation by omission, contrary to Hall. Here is the example:



Overdetermination by Omission

This diagram can be labelled a case of *Symmetric Overdetermination by Omission*, meaning that the absences of two events are equally responsible for an effect. Given that the omissions play exactly the same causal role as positive events do in cases of normal, positive, symmetric overdetermination, intuitively both omissions $\neg C$ and $\neg D$ should be judged as being causes of the effect E. Neither the Hall07 definition nor the BV12 definition respect this intuition.

Hitchcock (2009) gives the following informal story that could be modelled by this diagram.

Example 6. *A patient will die unless he receives two doses of medicine, which are currently in the custody of doctors C and D. Both doctors withhold the medicine, and the patient dies.*

Intuitively, most people agree that each doctor witholding his dose of medicine is a cause of the patient's death.

5.3.2 The Sufficiency of Counterfactual Dependence

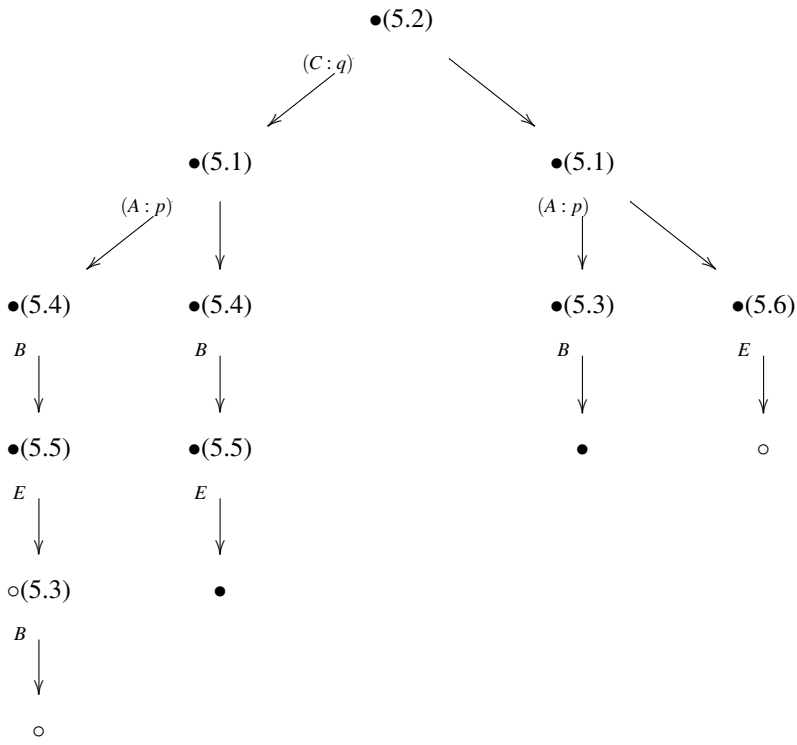
All definitions of causation considered in this work belong to the counterfactual tradition, which was initiated by Lewis (1973) (and even before that by Hume (1748)). Despite their differences, definitions in this tradition share the assumption that if E is counterfactually dependent on C then C is a cause of E. Given that most approaches only explicitly consider deterministic examples, we here take counterfactual dependence to be deterministic. (In other words, Definition 2 when restricted to cases where T^b is a deterministic theory.) Unfortunately one can construct examples to show that the BV12 definition violates this assumption.[1]

[1]Most authors that consider non-determinism still endorse the assumption even when dependence is generalised to non-deterministic dependence, but as the example shows we do not need this stronger assumption to make our point.

Consider the following theory, and one of its probability trees:

$(A : p) \leftarrow .$ (5.1) $B \leftarrow A.$ (5.3) $E \leftarrow B \wedge C.$ (5.5)

$(C : q) \leftarrow .$ (5.2) $B \leftarrow C.$ (5.4) $E \leftarrow \neg B \wedge \neg C.$ (5.6)



Now consider the story represented by the leftmost branch, where we have $\{C, A, B, E\}$, and – according to the BV12 definition – it is C that causes B by means of law (5.4). As in the *Late Preemption* example, the last edge represents the fact that A would have caused B, if B had not occurred yet. Looking at the relevant theories below, it is easy to verify that although E is counterfactually dependent on C, the BV12 definition does not judge C to be a cause of E.

Theory $T^b|do(\neg C)$ for Dependence

$A \leftarrow .$

$B \leftarrow A.$

$B \leftarrow C.$

$E \leftarrow B \wedge C.$

$E \leftarrow \neg B \wedge \neg C.$

Theory $T^*|do(\neg C)$ for BV12

$A \leftarrow .$

$B \leftarrow C.$

$E \leftarrow B \wedge C.$

$E \leftarrow \neg B \wedge \neg C.$

BV12 reaches this mistaken verdict because it considers the law $B \leftarrow A$ irrelevant: the effect (B) had already occurred by the time it was applied. Recall from Section 3.3 that dismissing laws that were applied after the effect as irrelevant proved essential in order for the BV12 definition to handle *Late Preemption*. Unfortunately this means there is no easy way the BV12 definition can be adapted in order to avoid this problem. (Note though that the example here presented is somewhat artificial. Specifically, it is not easy to come up with real-life situations in which the combination of laws (5.5) and (5.6) occurs.)

5.4 Conclusion

We have discussed several problems regarding the two main definitions of actual causation developed in Chapter 3.

First, the Hall07 definition turns out to be sensitive to details of the story which do nothing to change our intuitions about them. This sensitivity is an unwelcome side-effect of the fact that this definition distinguishes between cases of *Early Preemption* and cases like *Switch* without making explicit use of non-determinism. Since BV12 was designed from the outset using the non-deterministic language of CP-logic, it does not suffer from this problem.

Second, both definitions are unable to properly handle cases of *Symmetric Overdetermination by Omission*. The BV12 definition handles ordinary cases of *Symmetric Overdetermination*, i.e., cases where the causes are events, by exploiting the fact that both overdetermining causes are represented in separate laws. Overdetermination cases involving an omission lack this feature, which is why the BV12 definition fails.

Third, the BV12 definition fails to satisfy a basic assumption regarding causation, namely that counterfactual dependence is sufficient for causation. The reason for this failure lies in the method of handling *Late Preemption*: to capture the principle that causes come before effects, laws which were applied after the effect are dismissed entirely as irrelevant.

In Part II we develop a definition of causation using both a different language and a different methodology than we used in Part I. Instead of using CP-logic and a

general definition that builds on counterfactual dependency, we use structural equations extended with a timing and build up the definition by considering basic principles which it should satisfy.

As a result, the definition of actual causation from Part II is able to avoid all of the problems from this chapter. First, it distinguishes between *Early Preemption* and *Switch* using non-determinism. Second, events and omissions are treated more symmetrically than here (but still a little different, notably regarding their temporal properties). Third, rather than dismissing entire causal mechanisms which were temporally preempted, it keeps track of the timing of all events (and omissions).

Part II

A Principled Approach to Defining Actual Causation

Chapter 6

A Principled Approach to Defining Actual Causation

This chapter was previously published as:
Beckers, S., and Vennekens, J. (2016b). A principled approach to defining actual causation. *Synthese*, *forthcoming*.

In this chapter we present a new proposal for defining actual causation. We do so within the popular counterfactual tradition initiated by Lewis, which is characterised by attributing a fundamental role to counterfactual dependence. Unlike the currently prominent definitions, our approach proceeds from the ground up: we start from basic principles, and construct a definition of causation that satisfies them. We define the concepts of *counterfactual dependence* and *production*, and put forward principles such that dependence is an unnecessary but sufficient condition for causation, whereas production is an insufficient but necessary condition. The resulting definition of causation is a suitable compromise between dependence and production. Every principle is introduced by means of a paradigmatic example of causation. We illustrate some of the benefits of our approach with two examples that have spelled trouble for other accounts. We make all of this formally precise using structural equations, which we extend with a timing over all events.

6.1 Introduction

Although progress on the problem of actual causation has been made over the last decade, not a single definition on offer goes uncontested. In this chapter we develop a new proposal for defining actual causation. In comparison to the large number of

proposals out there, our approach offers the important benefit that it starts from basic principles. Indeed, many existing definitions lack proper foundations. Even when a detailed justification is given, it mostly consists of informal guidelines rather than precise formal conditions. By contrast we aim to make explicit what principles we take to be fundamental to causation, and show their consequences on particular examples. In this manner even those who disagree with the verdicts of our definition are guided to the principles from which they follow.

As a starting point, we delineate the borders of the search space we wish to explore. This implies formulating a sufficient and a necessary condition. The former serves as a lower bound, in the sense that its extension is a subset of all cases of actual causation, whereas the latter forms an upper bound. These conditions thus form the boundaries of a spectrum of concepts that contains actual causation somewhere in between. The task before us is to provide principles which point towards a single concept in this spectrum.

The literature on actual causation abounds in convoluted examples that discredit or confirm definitions of causation. To make matters worse, these definitions themselves often turn out to be quite hard to understand. To avoid these pitfalls we illustrate every principle by a very simple example, and indicate how the intuition behind it can be made formally precise using structural equations. To obtain maximal clarity, all but one of these examples are made up of the same ingredients, namely two protagonists named Billy and Suzy, each holding a rock in their hand, and a bottle that is standing a bit further waiting to be shattered. Hall and Paul (2003) introduced these types of examples, which are now widespread in the literature. Small changes to the details of the scenario suffice to highlight what we take to be the fundamental issues of the debate. Although we view it as a benefit of our approach that it can be developed using the simplest of examples, we also show how it handles two examples that have spelled trouble for other accounts.

In Part I we discussed a graded, probabilistic notion of causation, but here we restrict ourselves to a binary concept, i.e., we are purely interested in the question whether or not something is a cause. Further, for the most part we limit ourselves to deterministic examples. Also, we set aside the interesting recent research regarding the influence that norms and expectations have on our causal judgments that resulted in an extension to actual causation in Chapter 4. In the current part the focus is solely on the core problem of deciding whether or not something is an actual cause.

In the next section we present dependence as a sufficient condition for causation, followed by a necessary condition in Section 6.3. Sections 6.4 and 6.5 then present production as a necessary but insufficient condition lying in between the previous ones. Section 6.6 addresses important issues arising from non-determinism, including a comparison between the HP approach and ours. In Section 6.7 we refine our conditions by having a more detailed look at dependence, which narrows down the search space to a single option by discussing another example in Section 6.8. Section 6.9 interprets the resulting definition as a compromise between the concepts of *counterfactual dependence* and *production*. To conclude, Section 6.10 discusses two examples which

other definitions are unable to handle.
As we have done in Part I as well, throughout we take C and E to be endogenous literals, where C is a candidate cause for the effect E.

6.2 Counterfactual Dependence

Consider the first of our rock-throwing stories:

Example 7. *Suzy throws a rock at a bottle, while Billy idly stands by with a rock in his hand, having no intention to throw it. Suzy's rock hits the bottle, at which point it shatters.*

We need three endogenous variables to formally represent this story: BS represents the event that the bottle shatters, $Suzy$ and $Billy$ that Suzy, respectively Billy, throw a rock. We do not model the causes of them throwing, and just take this to be determined by the background conditions, i.e., the context. Thus an appropriate causal model M for this example consists of the single equation $BS := Billy \vee Suzy$.
This example is then represented by the context such that only Suzy throws her rock, resulting in the assignment $\{Suzy, \neg Billy, BS\}$. Without hesitation everyone would agree that Suzy throwing her rock caused the bottle to shatter. This judgement can be justified by a straightforward counterfactual observation: if Suzy had not thrown her rock, then the bottle would not have shattered. To formalise this, we present the structural equations version of the Dependence definition from Section . For the time being, we restrict attention to deterministic theories. Hence we get:

Definition 16. *Given a causal setting $(M, \vec{u})$ such that $(M, \vec{u}) \models C \wedge E$, E is* counterfactually dependent *on C if $(M_{do(\neg C)}, \vec{u}) \models \neg E$.*

In words, E is counterfactually dependent on C if intervening on the value of C, while holding the context fixed, results in $\neg E$. In the example it is easy to see that indeed BS counterfactually depends on $Suzy$, but not on $\neg Billy$.
This simple example, and the way we treat it, accounts for the majority of our everyday causal attributions. Hence it should come as no surprise that Hume (1748) defined actual causation – causation, for short – as counterfactual dependence – dependence, for short.[1] Following him, dependence is taken by many to be an important intuition underlying causation (Hall, 2007; Halpern, 2015a; Halpern & Pearl, 2005a; Hitchcock, 2001; Lewis, 1973; Pearl, 2000; Weslake, 2015; Woodward, 2003). In fact, all of these authors agree, as do we, that dependence is sufficient for causation.[2] Therefore this assumption serves as our starting point.

[1] Surprisingly in the same breath he formulated a different definition as well, known as the regularity account, which is also still influential.

[2] Halpern (2015a) discusses this for all of the HP-approaches, and Weslake (2015) does so regarding most of the others.

Principle 1 (**Dependence**). *If E is dependent on C in a causal setting $(M,\vec{u})$, then C is a cause of E w.r.t. $(M,\vec{u})$.*

6.3 Contributing Cause

While dependence is sufficient for causation, it is well-known not to be necessary. Indeed, *Symmetric Overdetermination* (SO) and *Preemption* – both *Late* (LP) and *Early* (EP) – are notorious counterexamples. In this section we compare Example 7 to SO, postponing LP and EP until later.

Example 8. *[Symmetric Overdetermination] Suzy and Billy both throw a rock at a bottle. Both rocks hit the bottle simultaneously, upon which it shatters. Either rock by itself would have sufficed to shatter the bottle.*

In terms of our causal model, this story represents the context in which both Suzy and Billy throw. Intuitively, most people judge each throw to be a cause of the bottle shattering. However it is easy to see that it is dependent on neither (although it is dependent on at least one rock being thrown, i.e., $M_{do(\neg Billy),\neg Suzy)},\vec{u}) \models \neg BS$). Despite the lack of dependence, there still is a sense in which we can legitimately say that each throw *contributed* to the shattering of the bottle.
To clarify this notion of contributing, let us zoom out for a second and consider the general causal model, rather than this specific story. At the general level, i.e., in absence of any information regarding the context $\vec{U}$, all we can say is that both *Suzy* and *Billy* could contribute to *BS*. Formally, we introduce the concept of a *contributing cause* to express this, which is also defined by Weslake (2015) and Woodward (2003).[3] First, we define the following helpful concept.

Definition 17. *We define that a consistent set of literals L is* sufficient *for a literal L_i w.r.t. M if $\bigwedge L \Rightarrow \phi_{L_i}$ and L_i is positive, or $\bigwedge L \Rightarrow \neg\phi_{L_i}$ and L_i is negative. Here, $\bigwedge L$ denotes the conjunction of all elements of L.)*

Recall that M consists of a set of equations of the form $V_i := \phi_{V_i}$, where ϕ_{V_i} is a propositional formula. Then, according to Definition 17, L is sufficient for L_i w.r.t. M just in case the conjunction of literals in L logically entails the propositional formula ϕ_{V_i} (when $L_i \equiv V_i$), or the propositional formula $\neg\phi_{V_i}$ (when $L_i \equiv \neg V_i$).
For example, in our rock-throwing model, $\{Suzy\}$ is sufficient for BS because $Suzy \Rightarrow Suzy \vee Billy$ is a logically valid implication, and $\{\neg Suzy, \neg Billy\}$ is sufficient for $\neg BS$ because $\neg Suzy \wedge \neg Billy \Rightarrow \neg(Suzy \vee Billy)$ is trivially valid.

Definition 18. *Given M, we define that C is a* direct possible contributing cause *of E if there exists a set of literals L containing C, such that L is sufficient for E, but $L \setminus \{C\}$ is not. We call L a* witness *for C w.r.t. E.*

[3] Our formulation and the ensuing principle are not entirely identical to theirs, but the difference is negligible.

Note that this definition is context-independent, and that only literals which appear in the equation for E can ever be direct possible contributing causes. To illustrate, both *Suzy* and *Billy* are direct possible contributing causes of *BS*, with witnesses $\{Suzy\}$ and $\{Billy\}$ respectively: $\{Suzy\}$ and $\{Billy\}$ are both sufficient for *BS*, and $\emptyset$ is not (since $\top \not\Rightarrow Suzy \vee Billy$).
More generally, the connection between two literals need not be direct:

Definition 19. *Given M, we define that C is a* possible contributing cause *of E if there exist literals $C = L_1, \ldots, L_n = E$ so that each L_i is a direct possible contributing cause of L_{i+1}.*

Besides the sufficiency of dependence, all authors mentioned earlier also agree on the principle that if C is an actual cause of E, then C has to be a possible contributing cause of E.[4] Indeed, if C is not a possible contributing cause of E, then under no circumstances does it affect the truth of E.
A natural step is to zoom in again, and refine this concept and its corresponding principle to the level of an actual storym by plugging a specific context $\vec{u}$ into the model M.

Definition 20. *Given $(M, \vec{u}) \models C \wedge E$, we define that C is a* direct actual contributing cause *of E if C is a direct possible contributing cause of E with a witness L such that $(M, \vec{u}) \models L$.*

Using this notion allows us to differentiate between the role of *Billy* in the contexts of Example 7 and Example 21. For Example 7, we have that $(M, \vec{u}) \not\models Billy$, and hence there is no witness for *Billy* being a direct actual contributing cause of *BS*. For Example 21, on the other hand, we have that $(M, \vec{u}) \models Billy$, and hence $\{Billy\}$ is a witness to the fact that *Billy* is a direct actual contributing cause of *BS*. Again we can generalize this concept by considering a chain of direct contributing causes.

Definition 21. *Given $(M, \vec{u}) \models C \wedge E$, we define that C is an* actual contributing cause *of E if there exist literals $C = L_1, \ldots, L_n = E$ so that each L_i is a direct actual contributing cause of L_{i+1}.*

From now on we speak simply of C contributing to E, rather than saying that C is an actual contributing cause of E. We now formulate our second principle, which provides a necessary condition for actual causation and therefore delineates the upper border of our spectrum.

Principle 2 (Contributing). *If C is a cause of E in a causal setting $(M, \vec{u})$, then C contributes to E w.r.t. $(M, \vec{u})$.*

Informally, what this principle states is that all actual causes of E are literals that contributed to satisfying/falsifying a formula ϕ_{V_i} for some variable V_i, which in

[4] For details regarding most of the approaches, again see (Weslake, 2015).

turn contributed to satisfying/falsifying another formula ϕ_{W_i}, etc., which in the end contributed to satisfying/falsifying ϕ_E.
The only difference between this principle and the one mentioned after definition 19, is that we have filled in an actual context. Weslake's definition (2015) has this principle directly built into it, as his (STRAND) condition. The reader may verify that all the other definitions mentioned above also satisfy the principle proposed here, as long as one takes into account the restriction to Boolean endogenous variables made earlier. Although this restriction is of no importance for the overwhelming majority of cases, there is one exception. In models that represent "trumping causation" by means of a three-valued variable, some of these definitions do call an event a cause even though it fails to contribute to the effect. However, the majority of authors agree that this is the wrong answer.[5] Hence if anything, this speaks in favour of **Contributing**.
Applying this principle allows us to exclude certain literals that clearly are not causes, such as $\neg Billy$ in Example 7. We thus now distinguish three relations between C and E:

- E is dependent on C. (*Suzy* in Example 7)
- C does not contribute to E. ($\neg Billy$ in Example 7)
- E is not dependent on C, but C does contribute to E. (*Suzy* and *Billy* in Example 21)

By **Dependence** and **Contributing** we know that there is causation in the first, and not in the second, of these cases. (That the cases are mutually exclusive, and thus the conjunction of our principles consistent, follows from Theorem 9 in Section 6.4.) Since *Suzy* and *Billy* are causes in Example 21, we might hope that besides being necessary, contributing is also sufficient for causation. In the following two sections we present two counterexamples to the sufficiency of contributing, and develop two principles which explain what may prevent contributing literals from being causes.

6.4 Production

The following story is paradigmatic in the literature for what has come to be known as *Late Preemption* (LP).

Example 9 (Late Preemption)**.** *Suzy and Billy both throw a rock at a bottle. Suzy's rock gets there first, shattering the bottle. However Billy's throw was also accurate, and would have shattered the bottle had it not been preempted by Suzy's throw.*

In this story, the process of Billy throwing a rock and shattering the bottle is *preempted* by the process involving Suzy, and this happens *after* the effect has occurred, i.e., after

[5] See (Weslake, 2015)[p.17] for a discussion.

the bottle has shattered. This is in contrast to *Early Preemption* (EP), in which a process is preempted before the effect occurs.[6] (See Section 6.6.)

As in SO, the bottle shattering is overdetermined by both throws, and again the bottle's state is not dependent on either throw. The difference here is the asymmetry that Suzy's rock hits the bottle, while Billy's does not. Our causal judgments reflect this asymmetry, as people unanimously judge Suzy's throw to be the sole cause.

How should we formally represent this example? If we continue using the three-variable causal model $BS := Suzy \vee Billy$, then we end up with the same causal setting as in SO. Since $Billy$ is a cause in SO, but not in LP, we need to refine our representation to take into account the difference between them. More specifically, we need to represent precisely that difference which justifies the shift in causal status of $Billy$ when going from SO to LP.

As noted, the difference between SO and LP is whether or not Billy's rock hits the bottle. Hence, one might distinguish between the two cases by adding variables SH and BH to the model, which represent Suzy's, respectively Billy's, rock hitting the bottle. Using such a model allows one to make the following observation: if we hold fixed BH at its actual value, then BS is dependent on $Suzy$ in case of LP, but not so in case of SO. This approach is taken by Halpern and Pearl (2005a) whose work on actual causation has set the benchmark for others to compare with. Their definition – in its many versions – takes full advantage of holding fixed variables at specific values regardless of their equations, given that certain structural criteria are fulfilled. (See Section 2.2.1 for the details. Recall that they refer to these non-standard causal settings as *structural contingencies*.)

We discuss the HP approach in Section 6.6.1, for now suffice it to say that we believe this approach is flawed, for it does not take into account the *reason why* Billy's rock did not hit the bottle, despite him throwing it. Yet that reason is obvious: Billy's rock arrived at the bottle *too late*. Or, in the words of Hall (2004)[p. 8]:

> Once the bottle has shattered, however, it cannot do so again; thus the shattering of the bottle prevents the process initiated by Billy's throw from itself resulting in a shattering.

If there is any principle regarding causation which is accepted across the board, then it is the fact that causes come before – or at most simultaneous with – effects. Our approach to handle LP consists in combining said principle with the temporal information regarding Billy's rock.

In order to formally represent this idea, it is necessary to represent the completion of each of the competing processes. The most obvious way to do so is by adding one variable for each process: SA represents Suzy's rock arriving at the bottle's location, and analogously for BA and Billy's rock. Our new causal model consists of the equations $BS := SA \vee BA$, $SA := Suzy$, and $BA := Billy$.

[6] These examples and this manner of distinguishing between them are due to Lewis (1986).

As with the original model, *Suzy* and *Billy* are possible contributing causes of *BS*. All we have done by adding the intermediate variables *SA* and *BA* is make explicit that the contributions of both *Suzy* and *Billy* to *BS* are mediated entirely through *SA* and *BA*, i.e., a thrown rock can only cause a bottle to shatter by flying through the bottle's location with sufficient momentum. Hence the question as to why *Billy* is not a cause of *BS* in LP is shifted to the same question regarding *BA*. The answer follows from a straightforward application of the accepted principle that causes come before effects, since *the bottle had already shattered by the time Billy's rock arrived*.

We will say of prevented processes and the associated literals, like *Billy* and *BA*, that they have been *preempted* for the effect. Literals that represent a process which completed succesfully, like *Suzy* and *SA*, will be referred to as *producers* of the effect. Given the essential role of temporal information, we choose to represent it separately from the variables. In this manner our approach is not dependent on there being suitable variables that capture the consequences of temporal asymmetry, like the variables *SH* and *BH* mentioned above. This representational clarity proves useful when dealing with cases of late preemption involving an omission, where such variables are unavailable and other approaches fail, as in Example 16 further on.

Definition 22 (Timing)**.** *A* timing τ *for a causal setting* $(M,\vec{u})$ *is a function from* $L_{(M,\vec{u})}$ *to* $\mathbb{N}$*. (Recall that* $L_{(M,\vec{u})}$ *is the set of all literals* L_i *such that* $(M,\vec{u}) \models L_i$*.)*

Informally, a timing can be interpreted as follows. An atom, like *Billy*, represents the fact that some event occurs in our story. Hence, if L_i is an atom, $\tau(L_i)$ simply represents the moment at which the event L_i happens, e.g., the moment that Billy throws his rock. If, on the other hand, L_i is a negated atom, like $\neg$*Billy*, then it represents an omission, i.e., it represents that some event does not occur. Since there is little sense in asking *when* an event does not occur, we take the pragmatic view that in this case $\tau(L_i)$ represents the moment at which the last event occurred that prevents $\neg L_i$ from happening, in the sense that the outcome of this event – together with the outcomes of all previous events – falsifies the formula ϕ_{L_i}.[7] Hence, the timing of omissions is derived from that of events.

We want to point out that aside from this temporal difference, we treat negated atoms and atoms symmetrically throughout this work, although some authors object to such a view.[8] This issue will pop up further on in the discussion of Example 20.

Also, by always interpreting atoms as events and negated atoms as omissions, the temporal asymmetry here introduced can be viewed as an implicit distinction between a *default* and a *deviant* value of a variable: only variables taking on their deviant value **true** have an independent timing, whereas the timing of variables remaining in their default value **false** is determined by the timing of the former. This perspective proves helpful when considering Example 16 further on. (We point out though that our version

[7] Here we are using the informal term "prevent" to get across the general idea. The precise interpretation of a timing is given in Definition 25.

[8] Halpern and Hitchcock (2015) provide some of the different views regarding this matter.

of the default/deviant distinction is rather minimal in comparison to other versions, such as for example that of Hitchcock (2007).)
If $\tau(L_i) < \tau(L_j)$, then this means that L_i happened/was prevented before L_j in the actual story. If $\tau(L_i) = \tau(L_j)$, then this means that both happened simultaneously, at least in the sense that the granularity of the story does not allow us to say which happened first.
Because not every story provides – or requires – complete temporal information, we also introduce the following concept.

Definition 23 (Partial timing)**.** *A* partial timing τ *for a causal setting* $(M,\vec{u})$ *is a partial function from* $L_{(M,\vec{u})}$ *to* $\mathbb{N}$.

Now that we have extended a story $(M,\vec{u})$ to include a timing, we can do the same for a counterfactual story $(M_{do(\neg C)},\vec{u})$: before $\neg C$, everything remains as it was in the actual story, after $\neg C$ the timing remains open.

Definition 24. *Given* $(M,\vec{u},\tau) \models C$*, we define* $\tau_{do(\neg C)}$ *as the partial timing that is identical to* τ *up until* $\tau(C)-1$*, has* $\tau_{do(\neg C)}(\neg C) = \tau(C)$*, and is not defined elsewhere.*

Because the structural equations represent causal relationships and causes must always precede their effects, the structure of the equations imposes restrictions on the timings that are possible. In particular, whenever V_i/$\neg V_i$ was caused at some time t, the causes that enabled/disabled ϕ_{V_i} must have already been present at this time. Further, as mentioned, an omission is caused at the same time as the last event which enabled it.

Definition 25. *Given* $(M,\vec{u},\tau)$*, for every n, we denote by* $L^n_{(M,\vec{u})}$ *the set* $\{L_i \in L_{(M,\vec{u})} \mid \tau(L_i) \leq n\}$*. For each endogenous variable* V_i *and the literal* L_i *containing* V_i *such that* $(M,\vec{u}) \models L_i$*, we define* τ *to be* valid *for* V_i *if*

- $L_i = V_i$ *and* $\tau(L_i) \geq \min_{k\in\mathbb{N}}\{L^k_{(M,\vec{u})}$ *is sufficient for* $L_i\}$*; or*
- $L_i = \neg V_i$ *and* $\tau(L_i) = \min_{k\in\mathbb{N}}\{L^k_{(M,\vec{u})}$ *is sufficient for* $L_i\}$.

A timing is valid for $(M,\vec{u})$ *if it is valid for all variables.*

For example, in our rock-throwing story where both Billy and Suzy throw, we require that $\tau(BS) \geq \tau(SA) \vee \tau(BS) \geq \tau(BA)$. In case Billy does not throw, we require that $\tau(\neg BA) = \tau(\neg Billy) = \tau(U_i)$, where U_i represents the exogenous event which prevents Billy from throwing. We can generalize the idea of validity to include partial timings.

Definition 26. *A partial timing* τ *is* possible *w.r.t.* $(M,\vec{u})$ *if there exists a timing* τ' *that extends* τ *(i.e.,* $\tau'(L_i) = \tau(L_i)$ *whenever* $\tau(L_i)$ *is defined) such that* τ' *is valid w.r.t.* $(M,\vec{u})$.

Using the timing, we can formalize the notion of production.

Definition 27. *Given $(M,\vec{u},\tau) \models C \wedge E$ with τ a valid timing for $(M,\vec{u})$, we define C to be a* direct producer *of E if C is a direct actual contributing cause of E w.r.t. $(M,\vec{u})$, with a witness L such that for each $L_i \in L$, $\tau(L_i) \leq \tau(E)$.*

More generally we define production in terms of a chain of direct producers.

Definition 28. *Given $(M,\vec{u},\tau)$ with τ a valid timing for $(M,\vec{u})$, we define C to be a* producer *of E if there exist literals $C = L_1,\ldots,L_n = E$ so that each L_i is a direct producer of L_{i+1}. For a partial timing τ', we define that C is a* producer *of E w.r.t. $(M,\vec{u},\tau')$ if there exists at least one valid timing τ that extends τ' such that C is a producer of E w.r.t. $(M,\vec{u},\tau)$.*

6.4.1 Comparison to Hall's Production

Our definition of production is inspired by that of Hall (2004), which we discussed in Section 3.4. There we expressed his definition in CP-logic as an instantiation of our general definition of causation, so that it could also incorporate negative literals as candidate causes. Still, our CP-logic version shares with Hall's original version the property that beyond the candidate cause C, all literals occurring in a chain of direct producers $C = L_1,\ldots,L_n = E$ will be positive.

The current definition of production includes all cases of production covered by these versions, but also allows the literals in a chain to be negative. For example, in Section 3.5 we presented *Double Prevention* as a paradigmatic case to illustrate how counterfactual dependence and production diverge: although there is dependence of G on C, according to the previous versions C does not produce G. We leave it to the reader to verify that according to the current definition of production C in fact does produce G. Examples 10 and 16 further on present two more illustrations.

As will become clear later, our more tolerant notion of production paves the way to a natural compromise between dependence and production into a single concept. We leave it to the reader to verify that, unlike the other two compromises from Part I, it does not suffer from any of the problems from Chapter 5.

Hall identified a problem with his definition of production, namely that it is context-sensitive. He illustrates this with the following example (Hall, 2004)[p. 31].

Example 10. *First imagine a scenario where we have $E := C \wedge D$, and both C and D are true. Then we zoom in on the details, and learn that the situation also involves an intermediate variable B, such that: $E := C \wedge \neg B$, and $B := C \wedge \neg D$.*

In both versions, E is dependent on both C and D, so according to our definition they are both causes of E, and thus also producers. According to Hall's definition of production, D is a producer of E in the first version only. Yet all the second version does is to make explicit some details that before were left implicit. In terms of the three original variables, the two models behave identically, namely E holds only if both of C and D do. In the second version, D prevents B, which would have prevented E,

making it a case of "double prevention". Because the chain from D to E contains an omission, it cannot fall under Hall's definition of production. From this he concludes that production must be context-sensitive, i.e., it depends on the level of detail that we use. Our definition of production, on the other hand, applies equally to both versions of the example. It therefore avoids Hall's relativistic conclusion.

6.4.2 Preempted Contributors

Producers are literals whose contribution helped bring about the effect. The following definition on the other hand generalizes the failure of Billy's contribution to do so.

Definition 29. *Given $(M,\vec{u},\tau) \models C \wedge E$, we define C to be* preempted *for E if C contributes to E w.r.t. $(M,\vec{u})$ and it is not a producer of E w.r.t. $(M,\vec{u},\tau)$.*

The difference between the role of Billy's throw in SO compared to LP, can now be expressed by saying that it changes from being a producer to being preempted. Concretely, any appropriate timing τ for LP will have $\tau(BA) > \tau(BS)$, whereas for SO, $\tau(BA) = \tau(SA) \leq \tau(BS)$. This allows us to exclude *Billy* from being a cause of *BS* in LP, by applying the formal counterpart of the aforementioned principle.

Principle 3 (**Preemption**). *If C is a cause of E w.r.t. $(M,\vec{u},\tau)$, then C is not preempted for E w.r.t. $(M,\vec{u},\tau)$.*

Combining **Contributing** and **Preemption** results in a stronger necessary condition for causation than **Contributing**:

Corollary 1 (**Producing**). *If C is a cause of E w.r.t. $(M,\vec{u},\tau)$, then C is a producer of E w.r.t. $(M,\vec{u},\tau)$.*

Extending the language of structural equations with explicit timings forms a substantial departure from existing structural equations approaches. However, one should not overestimate the role of a timing either. Looking at **Preemption** and Definition 29, we learn that the influence of a timing is limited to the timing of preempted events. Hence in practice it suffices to just give a partial timing over the literals that represent competing processes and their effect, such as *BA*, *SA* and *BS* in case of LP.
In all of the examples we have seen so far, producers were always causes. The next section shows that this is not necessarily the case.

6.5 Switches

Examples involving a switch make up another popular category to gauge intuitions on causation. In Section 5.2.2 we already discussed such an example in the form of a neuron diagram. The following is a story that usually accompanies that diagram (Hall, 2000, 2007; Halpern & Pearl, 2005a).

Example 11 (Switch)**.** *An engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the left-hand track, instead of the right. Since the tracks reconverge up ahead, the train arrives at its destination all the same.*

The following is an appropriate model for this story, where RT (LT) means that the train goes down the right-hand (left-hand) track, *Dest* means that the train arrives at its destination, and the context is such that *Switch* holds, i.e., the engineer flips the switch.

$$Dest := LT \vee RT.$$

$$LT := Switch.$$

$$RT := \neg Switch.$$

Intuitively most people agree that flipping the switch is not a cause for the train's arrival. But obviously it is a cause of the train going down the left-hand track, and this in turn is a cause of the train's arrival. Hence this is a counterexample to the transitivity of causation. Given that production is, by definition, transitive, it is also a counterexample to the sufficiency of production.

Part of the reason why we judge there to be no causation here is that the train would have arrived at its destination either way, i.e., there is no dependence. However we already know that dependence is not necessary for causation, so this is not the whole story. The further justification for our judgment is that the actual and the counterfactual story are too symmetrical in regards to the function of the switch. Flipping the switch contributes to a process that results in the train arriving. Not flipping the switch contributes to a different process, but one that has the exact same result. Therefore *Switch* and $\neg$*Switch* perform the same *causal role* in both stories, that of contributing to a process which results in *Dest*.

A fundamental property of causation, which underlies **Dependence** as well, is that causes are *difference makers*. Dependence expresses the strongest form of making a difference: to make a difference as to whether or not the effect takes place. What the switch example illustrates is that there is a weaker form of making a difference that is a necessary condition for causation, namely that the absence of a cause fulfills a different role than the cause itself. We formalize this property by means of the following principle.

Principle 4 (**Asymmetry**)**.** *If C is a cause of E w.r.t. $(M, \vec{u}, \tau)$, then $\neg C$ is not a cause of E w.r.t. $(M_{do(\neg C)}, \vec{u}, \tau_{do(\neg C)})$.*

This principle and the importance of difference making is defended as well by Sartorio (2005). Also Weslake (2015) incorporates a very similar principle into his definition of causation. However his formulation falls prey to a counterexample that we will discuss in Section 6.10.

Given the extreme symmetry between the actual story in Example 20 and the counterfactual story where the switch is not flipped, most definitions of causation will either judge both *Switch* and $\neg$*Switch* to be causes in their respective stories, or neither. Accepting **Asymmetry** delivers the intuitive verdict that neither is a cause. Note though that both *Switch* and $\neg$*Switch* are producers in their respective stories.
The qualification "most definitions" we made in the previous paragraph arises from the fact that some authors claim only events (and not omissions) can be causes, and therefore **Asymmetry** would be trivially satisfied for them. Recall that contrary to this view, we treat **true** and **false** symmetrically.
Before we move on, we need to address a possible objection. Some readers may not share our intuitions on the Switch example, on grounds that it is not at all certain the train will arrive either way. For instance, who is to say the right track would not break down? This is an important point, which can be made more vividly by using another famous counterexample to the necessity of dependence, namley *Early Preemption*. We direct our attention to this example, in order to show that we can do justice to these intuitions without dropping **Asymmetry**.

6.6 Non-determinism

Imagine yet another variant of our story:

Example 12. *[Early Preemption] Suzy throws a rock at a bottle. The rock hits it, and the bottle breaks. However Billy was watching Suzy, and would have thrown a rock just in case Suzy did not throw.*

This is an example of *Early Preemption*, because the causal mechanism connecting Billy's throw to the bottle shattering is preempted by Suzy already before the effect of the bottle shattering occurs. We can re-use the model from Example 9, except that the equation for *Billy* becomes $Billy := \neg Suzy$.
Only *Suzy* is directly dependent on the context $\vec{u}$, which is such that Suzy throws. Most authors consider examples of early preemption on a par with late preemption, and hence judge *Suzy* to be a cause of *BS* in this case as well. Yet if we compare this example to Switch, then we see that they are remarkably similar. *Suzy* plays the same role here as *Switch* does, namely it determines which of two processes occurs, where each process by itself is sufficient for the effect to take place.
Everything we just said about Switch, also holds for EP: in both cases the candidate cause – *Suzy* or *Switch* – is a producer of the effect (just as with SO and LP), and in the counterfactual situation the negation of the candidate cause is also a producer of the effect (contrary to SO and LP). In fact, the structural models and assignments to variables are almost completely identical in both cases.
But then how do we explain the prevalent intuition that *Suzy is* a cause of *BS* in EP? Here it becomes useful to consider the possibility that the effect will not occur either way. According to our current model, it is certain that if Suzy does not throw then the

bottle will shatter nonetheless. Surely that does not sound very realistic, as who is to say that Billy will not miss? All of our rock-throwing equations so far have assumed that Suzy or Billy throwing always results in the bottle's shattering. This assumption was a harmless simplification in the previous examples, because in each of them the *actual* story contained information on Billy's accuracy (with the exception of Example 7, where it was irrelevant). Because this is no longer the case here (since in the actual story Billy did not even throw), a proper analysis of EP requires incorporating this uncertainty. Hence we extend our model with variables *SAcc* and *BAcc*, representing Suzy and Billy's accuracy when they throw.

$$BS := SA \vee BA.$$

$$SA := Suzy \wedge SAcc.$$

$$BA := Billy \wedge BAcc.$$

$$Billy := \neg Suzy.$$

Allowing for the throws to be inaccurate changes the example significantly. In this chapter we have limited ourselves to deterministic examples, meaning we assumed that for each variable there was a definite truth-value in the actual story. The underlying motivation for this limitation is that as a result there is an unambiguous interpretation of the counterfactual story $(M_{do(\neg C)}, \vec{u})$, because that story corresponds to precisely one assignment of truth-values to all variables. To tackle *Early Preemption* we need to take a little excursion into the more general realm of non-deterministic examples, where there might be several counterfactual stories. Since there is no sense in which *BAcc* has a value in the actual story where Billy does not throw, we have to consider both the counterfactual story where Billy's rock hits the bottle and it shatters, and that in which he throws and misses.

Our approach can easily be generalised to allow for non-deterministic cases, by extending the context $\vec{U}$ with exogenous variables $\vec{W}$ whose values are undetermined in the actual story (eg., *BAcc* in EP).

Definition 30. *Given a causal model M over endogenous variables $\vec{V}$ and exogenous variables $\vec{U}$, we define a* partial context *as an assignment $\vec{u'}$ of values to variables so that $\vec{U'} \subseteq \vec{U}$, and refer to $(M, \vec{u'})$ as a* partial causal setting. *We call an assignment $\vec{w}$ to the remaining exogenous variables $W = U \setminus U'$ a* completion *of u'.*

Dependence is then defined as follows:

Definition 31. *Given a partial causal setting $(M, \vec{u'})$ such that for all completions $\vec{w}$ of $\vec{u'}$ we have: $(M, \vec{u'} \cup \vec{w}) \models C \wedge E$, E is* counterfactually dependent *on C if there exists a completion $\vec{w}$ such that: $(M_{do(\neg C)}, \vec{u'} \cup \vec{w}) \models \neg E$.*

All other definitions can be similarly generalised to partial causal settings. As before, actual causation is relative to a *story*. Up until now such stories have been represented

formally as a causal setting $(M, \vec{u})$. In the current more general setting, a story takes the form of a partial causal setting extended with a timing: $(M, \vec{u'}, \tau)$.
Our original Principle 4 is then replaced with:

Principle 4 (Asymmetry version 2). *If C is a cause of E w.r.t.* $(M, \vec{u'}, \tau)$*, then there exists a completion* $\vec{w}$ *of* $\vec{u'}$ *so that* $\neg C$ *is not a cause of E w.r.t.* $(M_{do(\neg C)}, \vec{u'} \cup \vec{w}, \tau_{do(\neg C)})$.

As a consequence, by adding the appropriate variables allowing for several counterfactual stories, we are able to do justice to our intuitions in both *Switch* and EP. If it is realistic to assume that train tracks do not malfunction, then the train will arrive either way and flipping the switch is not a cause. If on the other hand our intuitions do not support this assumption, then possibly the train would not arrive but for flipping the switch, and hence flipping it is a cause.
In the EP example, the counterpart of the malfunctioning track is Billy missing the bottle. Since it is quite plausible to take the accuracy of a boy throwing a rock to be much more uncertain than a sturdy track breaking, it is to be expected that intuitions for Suzy's throw being a cause are more common than those for flipping the switch. Hence an appropriate model for EP should contain a variable representing the uncertainty of the counterfactual story, contrary to a model for *Switch*. The more general non-deterministic version of **Asymmetry** then gives the right answer in both cases.
The lesson to be learned here is that structurally there is no difference between examples labelled "switches" and those commonly taken to exhibit early preemption. The difference lies in the reliability of the background process which might produce the effect in the absence of the actual process. Having expounded the importance of non-determinism in these examples, to keep things simple from here onwards we focus again on the deterministic version of **Asymmetry**.

6.6.1 Comparison to HP

As mentioned, for reasons of simplicity most structural equations approaches stick to deterministic models. Still, all of them claim to provide an adequate analysis of both *Early Preemption* and *Switch*. In Chapter 5 we already discussed the problems Hall's account faces in trying to distinguish between the two examples in a deterministic context. The BV12 definition, on the other hand, avoided these problems by using the non-deterministic solution here presented. To further justify our use of non-determinism, we again have a look at the HP approach. Halpern and Pearl (2005a) apply the same reasoning to *Switch* as we have:

> Is flipping the switch a legitimate cause of the train's arrival? Not in ideal situations, where all mechanisms work as specified. But this is not what causality (and causal modeling) are all about. Causal models earn their value in abnormal circumstances, created by structural contingencies, such as the possibility of a malfunctioning track. It is this possibility that should enter our mind whenever we decide to designate each track as a separate

> mechanism (i.e., equation) in the model and, keeping this contingency in mind, it should not be too odd to name the switch position a cause of the train arrival (or non-arrival).

Note that they explicitly refer to "the possibility of a malfunctioning track" as a structural contingency. On the face of it this suggests that the motivation behind their approach for dealing with Early Preemption/Switch is very similar to ours: if it is considered a significant possibility that the backup mechanism fails, then this possibility should be taken into account to assess causation. Concretely, in that case we should take into account the counterfactual story where the candidate cause does not occur, and the backup mechanism fails. Which factors determine whether or not the failure of the backup mechanism – be it a train track or a person throwing a rock – is a significant possibility is mostly an empirical matter, and should be decided on a case by case basis. We find further confirmation of our interpretation by considering another example of *Early Preemption*, discussed by Halpern and Pearl (2005a)[p. 30]. We present here the original formulation by McDermott (1995).

Example 13. *[Early Preemption 2] Suppose I reach out and catch a passing cricket ball. The next thing along in the ball's direction of motion was a solid brick wall. Beyond that was a window.*

Is catching the ball a cause of the window being safe? Even without giving a structural model to go along with this story, the similarity to Example 12 and *Switch* is obvious. Again, the answer depends on whether or not we consider the possibility that the backup mechanism – the wall blocking the window – will fail. Intuitively, most people judge this example to be more similar to *Switch* than to *Early Preemption*, meaning they do not judge catching the ball to be a cause. This is consistent with our approach: as with the failure of train tracks, people generally do not consider it a significant possibility that a solid brick wall will fail to stop a cricket ball. Halpern and Pearl (2005a) also treat this example similar to *Switch*:

> If we make both the wall and the fielder endogenous variables, then the fielder's catch is a cause of the window being safe, under the assumption that the fielder not catching the ball and the wall not being there is considered a reasonable scenario. On the other hand, if we take it [sic] for granted the wall's presence (either by making the wall an exogenous variable, not including it in the model, or not allowing situations where it doesn't block the ball if the fielder doesn't catch it), then the fielder's catch is not a cause of the window being safe. It would remain safe no matter what the fielder did, in any structural contingency.

The difference between their approach and ours lies in the method used to represent the failure of the backup mechanism.[9] We choose to do so in a very straightforward fashion:

[9]This difference is not limited to Halpern and Pearl. Collins (2000) and Hitchcock (2001) use the same terminology when discussing which counterfactual scenarios ought to be considered. For example, confronted

all possible stories are modelled as some partial causal setting $(M, \vec{u}, \tau)$. Hence we interpret the deterministic model for *Early Preemption* as stating that it is *impossible* for Billy to miss when he throws. If this statement is considered inappropriate, then one should use the non-deterministic model given above, i.e., one should add a variable that represents Billy's accuracy and consider the possibility that he misses.

Since Halpern and Pearl restrict themselves to deterministic models, the choice between these two models is not available to them. This explains why they seek recourse in structural contingencies, as they need some other method to consider stories beyond the ones allowed by a structural model.

One could take this to imply that the difference here is merely a matter of taste, depending on one's preferred method to represent uncertainty. This is far from the truth. Halpern and Pearl use structural contingencies in a wide variety of cases, and these go well beyond examples resembling *Early Preemption*.

The criteria for deciding if a structural contingency may be used, conditions 1 and 2 in Definition 1, are not founded on underlying principles or heuristics that guide their application. As a result, they allow for a plethora of situations for which it is hard to see why we should consider them at all.[10] Indeed, Halpern and Pearl (2005a)[p.24] concede that in some cases their definition offers acceptable answers only if one explicitly stipulates which situations are "allowable settings". Therefore the interpretation of structural contingencies we have just given only applies to a limited number of cases.

To illustrate, we briefly return to *Late Preemption*. HP use the following model for this example, where *SH* and *BH* represent Suzy's, respectively Billy's, rock hitting the bottle:

$$BS := SH \vee BH.$$

$$SH := Suzy.$$

$$BH := Billy \wedge \neg SH.$$

We first have a look at whether or not this model is appropriate to capture the causal structure behind *Late Preemption*.

A first problem with this model is that the asymmetry between Suzy's throw and that of Billy is built right into the model: it does not allow for the story in which Billy throws faster than Suzy, or the story in which they both throw equally fast, as in

with Example 13, Collins (2000)[p. 8] says that "It is more far-fetched, on the other hand, to suppose that the brick wall be absent, or that the ball would miraculously pass straight through it." Considering an example involving a boulder – Example 18 given in the next chapter – Hitchcock (2001)[p. 298] says of the failure of the backup mechanism that "This possibility is just too far-fetched." Hall and Paul (2003)[p. 26] criticise Hitchcock by pointing out the arbitrariness in his use of this terminology.

[10] For details on these situations and the counterexamples they allow, see for example (Hall, 2007; Weslake, 2015). Halpern (2015a) has recently proposed a new definition which is more restrictive, avoiding some of these pitfalls, but not all. Further, it allows for new counterexamples, eg., it fails to judge each of *Suzy* and *Billy* a cause in case of SO.

Symmetric Overdetermination. There is nothing in the informal story in Example 9 to suggest that the difference in speed is a general, structural property. On the contrary, it sounds natural to assume that this difference is a contingent property of the actual story. However, as pointed out by Halpern (2015a), this problem can be set straight by also including *BH* into the equation for *SH*, and adding an exogenous variable to represent the order by which the rocks arrive. Hence this problem is of little consequence.

A second, more fundamental, problem, is the presence of *SH* in the equation for *BH*. Recall that a structural equation represents a causal mechanism, in this case the mechanism connecting Billy's throw to Billy's rock hitting the bottle. That mechanism does not involve *SH*, since Suzy and her rock form an entirely different and independent mechanism. Therefore, it seems conceptually wrong to include *SH* in the equation for *BH*. A consequence of this conceptual error is that if we consider the context where only Billy throws, then $\neg$*Suzy* is actually a producer (according to our definition) of *BS*. This is a very counterintuitive result.

The role played by $\neg$*SH* in the equation for *BH* is not that of a contributor to *BH*, but rather that of a constraint: it is supposed to capture the property that a bottle cannot shatter if it has already done so. This confirms that one cannot adequately deal with *Late Preemption* without making vital use of temporal information, as we argued in Section 6.4. Since HP stick to structural equations proper, they are forced to build this temporal information into the model itself. More specifically, the presence of $\neg$*SH* in the equation for *BH* compensates for the fact that they do not use a timing. Given these counterintuitive consequences, we prefer to use our symmetric model, containing the variables *SA* and *BA*.

Here it is useful to point out that for every approach using structural equations, the verdict given by a definition of causation is to a large degree dependent on the particular model being used. Since in many cases there is room for debate as to which model is appropriate for a given informal story, this means one can often counteract undesired outcomes of applying a definition by calling into question the model being used. (See (Halpern & Hitchcock, 2010) for a discussion of this issue.) However because our approach ultimately relies on basic principles, rather than on the intutiveness of examples, we believe it is less affected by this issue. If one accepts our principles, then one can make judgments about a causal model regardless of which informal story it is supposed to capture. In this manner, the problem of model appropriateness can to some extent be separated from the problem of defining actual causation.

Setting aside our disagreement regarding the choice of model for *Late Preemption*, we now turn to the HP approach and how it applies given their preferred model. It considers the structural contingency that Billy throws and yet fails to hit the bottle, even though Suzy does not throw. Contrary to the interpretation used for *Switch*, this structural contingency cannot be interpreted simply as the possibility that the backup mechanism fails to function properly, because the actual story explicitly stipulates that it does not: "Billy's throw was also accurate, and would have shattered the bottle had it not been preempted by Suzy's throw." This stipulation is not just a detail occurring in

our version of the example, but forms an essential part of *Late Preemption* cases.
One might object that there is also another possible interpretation, consistent with what has been said: namely that a structural contingency represents what is *generally possible*, rather than what is *possible given the actual story*. On this reading, the actual information that Billy was accurate is of no interest, all that matters is whether or not in general it is possible that he is not accurate. But going down this road leads to a slippery slope, for it blurs the distinction between *general* and *actual* causation. More specifically, if one can ignore the actual state of Billy's accuracy, then why not ignore other aspects of the actual story as well? For example, why not then consider the story in which Suzy throws but misses and Billy does not throw, and use it to conclude that Billy's throw also caused the bottle to shatter in *Late Preemption*?
Obviously according to the HP definition it is not the case that anything goes. Only those structural contingencies satisfying the conditions from Definition 1 may be considered. But what should be clear by now, is that it is not easy to come up with a consistent and systematic interpretation of what these structural contingencies are supposed to represent. Therefore we prefer to stay far away from them, and instead simply use a non-deterministic model to represent aspects of the story which are not actually determined, and use a timing to exclude those events which happened too late.

6.7 Dependence Revisited

So far, we have established that dependence is sufficient for causation but not necessary, while production is necessary but not sufficient. Therefore causation must lie in between these two concepts. To pinpoint its location, we present a theorem that relates dependence to production.

Theorem 9. *Given a valid timing τ, E is dependent on C w.r.t. $(M,\vec{u})$ if and only if both of the following conditions hold:*

- *[Condition 1]: C is a producer of E w.r.t. $(M,\vec{u},\tau)$.*
- *[Condition 2]: $\neg C$ is a producer of $\neg E$ w.r.t. $(M_{do(\neg C)},\vec{u},\tau_{do(\neg C)})$.*

Proof. The implication from right to left is trivial, hence we only need to prove the implication from left to right.
Assume E is dependent on C w.r.t. $(M,\vec{u})$, or in other words, $(M,\vec{u}) \models C \wedge E$ and $(M_{do(\neg C)},\vec{u}) \models \neg E$.
Take τ to be any valid timing w.r.t. $(M,\vec{u})$, $n = \tau(E)$, and $m = \min_{k\in\mathbb{N}}\{L^k_{(M,\vec{u})}$ is sufficient for $E\}$. We first prove that C is a producer of E w.r.t. $(M,\vec{u},\tau)$.
Take $L^1 \subseteq L^m_{(M,\vec{u})}$ to be minimally sufficient for E, i.e., L^1 is sufficient for E, and for any $L_i \in L^1$, $L^1 \setminus \{L_i\}$ is not sufficient for E. (Such a set can be constructed by removing elements from $L^m_{(M,\vec{u})}$ one by one.) By construction, all literals in L^1 are direct actual

contributors to E. Moreover, since $m \leq n$, these literals are direct producers of E as well.
Since $\vec{U} = \vec{u} \subset L_{(M_{do(\neg C)},\vec{u})}$, it follows that if $(L^1 \setminus \vec{U} = \vec{u}) \subseteq L_{(M_{do(\neg C)},\vec{u})}$, then $E \in L_{(M_{do(\neg C)},\vec{u})}$, i.e., $(M_{do(\neg C)},\vec{u}) \models E$. Therefore there exists at least one endogenous literal $D \in L^1$ such that $D \notin L_{(M_{do(\neg C)},\vec{u})}$. By the previous paragraph, D is a direct producer of E.
If $D = C$, then we are finished with this part of the proof. So assume $D \neq C$. We can apply the exact same reasoning as we did for E, to find a direct producer F of D such that $F \notin L_{(M_{do(\neg C)},\vec{u})}$. Since production is transitive, F is a producer of E as well. Given that there are only a finite number of endogenous literals, and that M is assumed to be acyclical, continuing this reasoning will eventually end up with finding C as a producer of E. Therefore we conclude that C is a producer of E w.r.t. $(M,\vec{u},\tau)$.
We can apply the exact same reasoning to prove that also $\neg C$ is a producer of $\neg E$ w.r.t. $(M_{do(\neg C)},\vec{u},\tau_{do(\neg C)})$, which concludes the proof.

□

Because this theorem indiscriminately applies to all valid timings, the first conclusion we can draw from it is that all information contained in a particular timing is lost when we consider dependence. Since we introduced the notion of a timing precisely to distinguish between cases where dependence was too crude a tool, this should not come as a surprise. On the contrary, the lesson learned from comparing examples such as *Symmetric Overdetermination* and *Late Preemption* was that the actual timing should be taken into account in order to judge actual causation. This theorem shows that without loss of generality, we can restrict our attention to one particular timing when comparing dependence and production. We now consider how the conjunction of Conditions 1 and 2 can be weakened, so that we shift from dependence to causation.
By **Producing**, we know Condition 1 should stay. Yet as the *Switch* example has shown, production does not satisfy **Asymmetry**: both *Switch* and $\neg$*Switch* are producers of *Dest* in their respective stories. Therefore a straightforward and natural suggestion is to combine production (Condition 1) with the constraint that **Asymmetry** should be satisfied. In other words, Condition 2 should be replaced with Condition 2': $\neg C$ is not a producer of E w.r.t. $(M_{do(\neg C)},\vec{u},\tau_{do(\neg C)})$.
Since *Switch* was the only example discussed which required us to look beyond production, it is easy to see that defining causation as the conjunction of Conditions 1 and 2' agrees with our judgments on all examples discussed so far. Note however that temporal information plays no role in judging *Switch*: the model is such that each story only allows one valid timing, and hence in this case the notions of producing and contributing are equivalent. Therefore we cannot rule out a slightly stronger alternative to Condition 2, let us call it Condition 2", where producing is replaced with contributing: $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)},\vec{u})$.

To decide between these two conditions, we now present an example in favour of adopting Condition 2', instead of Condition 2". In the next chapter (in Section 7.5.3) we offer a more principled argument in defence of Condition 2'.

6.8 Not Contributing vs. Not Producing

In this section we present a counterexample to the necessity of Condition 2", resulting in the acceptance of Condition 2'. However the example is rather exotic, since it is hard to even find examples for which these two options disagree. (We have not found any in the literature.) Hence we do not put much weight on our preference of Condition 2' over Condition 2", as in practice this will hardly ever matter.

Example 14. *In general, Billy throws rocks at bottles either if Suzy does not, or if he just feels like it. Today, Billy throws a rock at a bottle because he feels like it. Immediately afterwards Suzy throws a rock as well. Suzy's rock was thrown harder, and gets there first, shattering the bottle. However Billy's throw was also accurate, and would have shattered the bottle had it not been for Suzy.*

To model this variant of the rock-throwing story, which combines elements of early and late preemption, we need to adjust the equation for *Billy*, giving:

$$BS := SA \vee BA.$$

$$SA := Suzy \wedge SAcc.$$

$$BA := Billy \wedge BAcc.$$

$$Billy := Feels \vee \neg Suzy.$$

Here *Feels* means that Billy just feels like throwing, regardless of what Suzy does. Hence the context is such that *Feels* and *Suzy* hold. An appropriate timing τ is such that $\tau(Feels) \leq \tau(Billy) < \tau(Suzy) \leq \tau(SA) \leq \tau(BS) < \tau(BA)$. The question is whether or not *Suzy* is a cause of *BS*.

Given that Suzy's throw preempted Billy's throw from shattering the bottle, the example looks similar to LP, which suggests that *Suzy* is a cause. On the other hand, in the counterfactual story $do(\neg Suzy)$, Suzy's not throwing contributes to the process that would have Billy's rock shattering the bottle, just as with EP. Even more, we know that Billy was accurate, so there is no counterfactual story in which the bottle does not shatter, contrary to EP. Therefore the example is also similar to a switch, which suggests that *Suzy* is not a cause.

We believe the first similarity, to LP, to be the more fundamental one: even though it may hold in general that $\neg Suzy$ can produce *Billy*, in this story we already know that Suzy threw after Billy did. So in this case, Suzy throwing or not throwing was completely irrelevant to Billy's throw, which was instead produced by the fact that he

felt like throwing. Therefore when considering what would have happened if Suzy had not thrown, the right answer is that $\neg Suzy$ would not have produced anything (except for $\neg SA$), and just as with LP *Suzy* should be judged a cause of *BS*.
Now we compare how Condition 2" and 2' deal with this example. It is clear that *Suzy* produced *BS* in the actual story. In the counterfactual story, $\neg Suzy$ contributes to *BS*. Therefore this is a counterexample to the necessity of Condition 2".
The partial timing $\tau_{do(\neg Suzy)}$ has $\tau_{do(\neg Suzy)}(Feels) \leq \tau_{do(\neg Suzy)}(Billy) < \tau_{do(\neg Suzy)}(\neg Suzy)$. Therefore $\neg Suzy$ does not produce *Billy* w.r.t. $(M_{do(\neg Suzy)}, u, \tau_{do(\neg Suzy)})$, which implies it does not produce *BS* either. This is in agreement with the necessity of Condition 2'. We conclude from this that the right choice to make is to take the conjunction of Conditions 1 and 2' as a sufficient and necessary condition for causation.

6.9 Discussion and Results

Our principles have led us to propose the following definition of actual causation.

Definition 32. *[Actual Causation] Given* $(M, \vec{u}, \tau) \models C \wedge E$*, we define C to be an* actual cause *of E w.r.t.* $(M, \vec{u}, \tau)$ *if C produces E w.r.t.* $(M, \vec{u}, \tau)$ *and* $\neg C$ *does not produce E w.r.t.* $(M_{do(\neg C)}, \vec{u}, \tau_{do(\neg C)})$.

The precise formulation of this definition is dependent on the fact that we defined production over a partial timing as being a producer in at least one valid timing that extends it (as opposed to being a producer in all of them). This boils down to assuming that the default is for actual contribution to imply production, which is in line with our earlier observation regarding the limited influence of a timing: unless we know that a contributing process was preempted, it is a producer. As with the difference discussed in the previous section however, there are very few examples where this distinction matters.
Our definition of actual causation is built up entirely out of production, a concept which has so far received too little attention in the literature. A key property of production is that it focusses solely on the actual world: unsatisfied literals are entirely irrelevant. It tells us whether some event *brought forth* another as things actually happened.[11]
Causation shares production's interest in the actual world, but extends it with a contrast to a counterfactual world: did some event bring forth another as things actually happened, and if so, would the absence of said event not have brought forth the other? In the overwhelming majority of cases, if the first question is answered in the affirmative, so is the second; only examples exhibiting switching behaviour form an exception. This seems to suggest that the intense focus on the counterfactual nature of causation that we have observed in recent years is somewhat misguided. However

[11] In this respect it is similar to the notion of responsibility as it figures in ethics: ethical judgments concern (for the most part at least) what did happen, not what could have happened. We intend to examine this similarity in more detail in future work.

when we take into consideration Theorem 9, the picture becomes more nuanced, since dependence and production are tightly connected as well.

The main distinguishing feature of dependence is that it cares only about end results: it considers only whether C and E hold in the actual and counterfactual story, without looking at the temporal details – i.e., the timings – of how this came about. The importance of dependence therefore lies in its simplicity: one can forget about the intricacies of timing and preemption, and still end up with an answer that does the job most of the time.

We can express the difference between production, dependence, and causation in a nutshell by saying that production answers the "How?" question, dependence answers the "What if?" question, and causation answers the "Why?" question. The first is usually associated with understanding, the second concerns a form of *a posteriori* prediction, and the third is fundamental to explanation.

Although we have built up our definition using formal principles and theoretical examples, there has been empirical validation recently that points in a very similar direction. The idea that causation is a combination of dependence and production has been confirmed experimentally on a set of physical test-cases by Gerstenberg, Goodman, Lagnado, and Tenenbaum (2015), although their notion of production is less formal and somewhat different from ours. They too stress the importance of distinguishing between different ways a cause can make a difference to the effect (Gerstenberg et al., 2015)[p. 1]:

> We argue that the core notion that underlies people's causal judgments is that of difference-making. However, there are several ways in which a cause can make a difference to the effect. It can make a difference to *whether* the effect occurred, and it can make a difference to *how* the effect occurred.

We now have a look at our definition in practice by confronting it with some troublesome examples.

6.10 Some Examples

Weslake (2015) gives an overview of the most prominent definitions of actual causation in the structural equations framework. After presenting counterexamples to all of them, he proposes a definition that succeeds in getting the right answer for these examples. We leave it to the reader to verify that our definition delivers the correct verdict in these cases as well.[12] More interesting are his "non-structural counterexamples". These exhibit structural patterns that are identical to cases of symmetric overdetermination and early preemption, yet seem to give rise to different intuitions. He leaves it as an

[12] One should take into account our discussion of early preemption from Section 6.6 though: Weslake uses the deterministic model for EP and still judges there to be causation, whereas we claim there is causation only when using the non-deterministic model.

unsolved problem how to deal with these examples correctly as well.[13] Therefore we consider them as suitable test-cases for our approach.
The first example, named "Careful Poisoning", has the same structure as early preemption (Weslake, 2015)[p. 22].

Example 15. *Assistant Bodyguard puts a harmless antidote in Victim's coffee (A). Buddy then poisons the coffee (B), using a poison that is normally lethal, but which is countered by the antidote. Buddy would not have poisoned the coffee if Assistant had not administered the antidote first. Victim drinks the coffee and survives ($\neg D$).*

Intuitively, most people – but not all – agree that adding the antidote is not a cause of Victim's survival. Rather, it seems as if Assistant Bodyguard and Buddy are playing a trick on Victim: "we might suppose that Assistant Bodyguard is up for a promotion ... and wants to make it look as though he has foiled an assassination attempt. Buddy is helping him." (Hitchcock, 2007)[p. 520]. The model for this example is $D := \neg A \wedge B$, $B := A$, and the context is such that A holds. It is easy to see that A produces $\neg D$ in the actual story, and that $\neg A$ would likewise produce $\neg D$ in the counterfactual story. This example is thus nothing but a switch, and hence our definition does not consider A a cause of $\neg D$. Looking back at our discussion in Section 6.6, it is revealing that Weslake – and others with him – judges this example to be similar to Early Preemption, but fails to note the similarity to Switch. We can accommodate for the observation that some people have different intuitions here in the same manner as we did for those examples by pointing out that the backup process is assumed to be completely reliable, which might strike some as unrealistic.
The second example, named "Careful Antidote", is similar in structure to Examples 21 and 9 (Weslake, 2015)[p. 20].[14]

Example 16. *Assassin is in possession of a lethal poison, but has a last-minute change of heart and refrains from putting it in Victim's coffee ($\neg A$). Bodyguard puts antidote in the coffee (B), which would have neutralized the poison. Victim drinks the coffee and survives ($\neg D$).*

As with the previous example, adding the antidote intuitively is not a cause of Victim's survival. Once more this spells trouble for many definitions, given the resemblance to symmetric overdetermination, where our intuitions are reversed. We are able to look beyond this resemblance and handle this example as a case of *Late Preemption*, in the

[13] As a notable exception, Hall's account (2007) is able to deal with all of these examples succesfully. (Although he would have to add an extra variable to the model for the Backup example, and he disagrees with Weslake on the trumping causation example). Unfortunately it falls victim to other counterexamples, the most well-known being those from Hitchcock (2009). Again we leave it to the reader to verify that our definition does deliver the right verdict in all of the examples discussed there as well.

[14] An almost identical example is given by Hall (2007), named "back-up threat canceller". He uses it as an example that escapes his earlier dual-concept view of causation as being either dependence or production, and motivated him to develop his later definition. As the analysis will show, our more tolerant notion of production does capture this example. Thus it serves as a good illustration of how our notion of production extends his.

same manner as we distinguished between LP and SO, namely by using the timing. Recall that other approaches avoid having an explicit timing by adding additional variables, such as *SH* and *BH* in LP. Here, there are no obvious candidates for such variables, which explains why they struggle with this example.

Weslake uses the single equation model $D := A \wedge \neg B$. While our approach also gives the correct result for this model, we will explain our reasoning with the following more detailed one: $D := Dr \wedge L$ represents the fact that Victim dies if he drinks a lethal coffee, and $L := A \wedge \neg B$ represents the fact that the coffee is lethal if Assassin poisons it and Bodyguard does not add an antidote. The context is such that Dr, $\neg A$ and B hold. As $\neg D$ is dependent on $\neg L$, it is clear that $\neg L$ causes $\neg D$. Note also that any causal status of either Assassin or Bodyguard is mediated entirely through L. Since Assassin comes first, from the moment he refrains from putting poison in the coffee, it is not lethal. (Or to be precise, the coffee is no longer potentially lethal, as of course it was not lethal to start with.) Concretely, $\tau(\neg L) = \tau(\neg A) < \tau(B)$. Hence whatever Bodyguard's action might be, it is too late and is not a cause of $\neg D$, in agreement with our intuition. On the other hand if we change the story so that the order of Assassin and Bodyguard is reversed, then our definition *would* judge B a cause. Indeed, as soon as B happens, i.e., the antidote is added, the coffee has become poison-proof, i.e., no longer potentially lethal. Given that Assassin's omitting the poison happens afterwards, we see that B must be the cause of $\neg L$. Hence it is a cause of $\neg D$ as well.

Lastly, in case we are unable to tell which happened first, $\neg A$ or B, we call both of them causes, just as we did for SO.

To some calling B a cause, even when it happens first, may initially sound counterintuitive (while others may not have any clear intuitions here at all). Given the structural similarities between examples with conflicting, or simply confusing, intuitions, it is too much to expect of any definition that it will align perfectly with intuition in all cases. However an important benefit of our principled account, is that precisely by pointing out the similarities we can show how the same principles are at work in intuitively different examples, and possibly transform people's initial unreflective intuitions into informed judgments.

For example, it could be objected that according to our definition, even if in the end Victim changes his mind, and *does not drink the coffee*, B would nonetheless be a cause of $\neg D$.[15] There is no escaping the fact that initially this sounds counterintuitive. We believe the problem lies with the vagueness surrounding the nature of omissions, and their connection to time.[16] In this example, the omission is the statement "Victim does not die", or perhaps better, "Victim does not die from drinking poisonous coffee". At the start of the example Victim's death had not yet been prevented, and at the end it has. Hence there must be some particular event that happened in between such that Victim's death was prevented precisely at the moment this event occurred. The question

[15] For a very similar example, see "non-existent threats" (Hall, 2007).

[16] Hitchcock (2007) analyses these types of examples using a default/deviant distinction As we mentioned in Section 6.4, the temporal asymmetry between events and omissions in our notion of a timing can also be interpreted as invoking such a distinction.

is which event? Intuitions seem to be at a loss here, as there is no obvious candidate which presents itself. Certainly refusing to drink a perfectly fine coffee cannot be a cause of Victim's failing to die. We suggest that the way out is by using our principled approach, which generalizes the lessons learned from other examples about which we do have firm intuitions. Therefore the first event which prevented victim's death should be judged its cause.

We come back to the original story to illustrate the vagueness regarding omissions and their timing. The story states the omission that Assassin does not put the poison in Victim's coffee, and that he does not do so because he has a last-minute change of heart. The fact that the statement regarding Bodyguard follows the one regarding Assassin, indicates a temporal order: first, Assassin refrains from putting in poison, then, Bodyguard adds antidote. But intuitively it is not at all clear what it means for Assassin's omission to occur first, precisely because it is not clear what event occurs (and when it does so) such that Assassin's mental state shifts from "intending to put poison in Victim's coffee" to "no longer intending to put poison in Victim's coffee".

This is confirmed if we adapt the example so that we focus only on the timings of events, as with *Late Preemption* and *Symmetric Overdetermination*. Imagine that we start out with a coffee that is already poisonous, and both Assassin and Bodyguard add an effective antidote. In that case our intuitions would simply follow the temporal order by which the antidotes were added: if Assassin adds his first, then Bodyguard adding the antidote is not a cause, and vice versa. We claim that, by analogy, it makes sense to call Assassin's refusal to poison the coffee a cause, as long as he makes up his mind *before* the antidote is put in.

6.11 Conclusion

Our goal in this chapter has been to construct a definition of actual causation from the ground up. We have formulated several principles which we take to be fundamental properties of causation, and illustrated each of them by way of a simple example. As a result we derived a definition that is a compromise between the pull of two distinct concepts, namely dependence and production. Given that all three concepts agree on a large number of examples, it is not surprising that the distinction between them is often neglected or misunderstood.

We have applied our definition successfully on a number of paradigmatic examples: symmetric overdetermination, late/early preemption, switching, careful poisoning, and careful antidote. In addition, we have also checked our definition against all examples found in (Hall, 2000, 2004, 2007; Halpern, 2015a; Halpern & Pearl, 2005a; Hitchcock, 2001, 2007, 2009; Weslake, 2015) and many more. Our definition can be applied to all of them along the same lines as we have applied it to the examples mentioned.

We hope our principled approach proves useful as well to those who contest our resulting definition, by clarifying formally how causal judgments depend on accepting

or rejecting the underlying principles. Further, we believe the interplay between the three concepts here described offers a fruitful perspective for understanding different aspects and interests present in causal stories. In future work we intend to apply this insight by comparing the role of causation in different domains, such as the positive sciences, history, and ethics.

Lastly, our principled definition makes it easier to argue for or against specific causal judgments regarding complex examples. Despite the fact that our definition agrees with intuition in simple paradigmatic cases, we are not forced to seek recourse in intuitions to justify our answers in all cases. Given the diversity of intuitions and their mutual inconsistency, it is essential to have a principled method to settle causal judgments one way or the other.

Chapter 7

The Transitivity and Asymmetry of Actual Causation

The counterfactual tradition to defining actual causation has come a long way since Lewis started it off. However there are still important open problems that need to be solved. One of them is the (in)transitivity of causation. Endorsing transitivity was a major source of trouble for the approach taken by Lewis, which is why currently most approaches reject it. But transitivity has never lost its appeal, and there is a large literature devoted to understanding why this is so. Starting from a survey of this work, we will develop a formal analysis of transitivity and the problems it poses for causation. This analysis provides us with a sufficient condition for causation to be transitive, a sufficient condition for dependence to be necessary for causation, and several characterisations of the transitivity of dependence. Finally, we show how this analysis leads naturally to several conditions a definition of causation should satisfy. Our own definition developed in the previous chapter does indeed satisfy these conditions, hence this chapter offers further support for it.

7.1 Introduction

As we saw in the previous chapter, all approaches under consideration take as their starting point the assumption that counterfactual dependence is sufficient for causation, but not necessary (Hall, 2004, 2007; Halpern, 2015a; Halpern & Pearl, 2005a; Hitchcock, 2001; Lewis, 1973; Weslake, 2015; Woodward, 2003). That dependence is sufficient is usually accepted simply as a fundamental principle underlying causation. That it is not necessary, on the other hand, is usually defended by pointing to intuitively

strong counterexamples. Lewis (1973) forms an important exception to this rule, as he defends the lack of necessity by invoking a principle as well, namely that causation is transitive: causation is transitive, dependence is not, therefore there can be causation without dependence.

The first strategy, that of offering counterexamples, has proven most successful. There are two reasons for this. First, almost everyone besides Lewis rejects the transitivity of causation. Second, there are counterexamples to the necessity of dependence that have nothing to do with transitivity. Despite its success, this strategy has to date not offered a general insight into precisely when or why the transitivity of causation breaks down. Although a substantial number of authors have addressed the problem of transitivity, none of them offers a generally sufficient condition for causation to be transitive (Hall, 2000, 2004; Halpern, 2015b; Halpern & Pearl, 2005a; Hitchcock, 2001; McDermott, 1995; Paul & Hall, 2013; Sartorio, 2005). A recent discussion by Halpern (2015b) does formulate several sufficient and necessary conditions for transitivity, however those apply only to cases where there is dependence.

The main contribution of this chapter is to offer a principled explanation of why transitivity should be rejected as a general condition, while also offering conditions under which it should be satisfied. Specifically, we will explain both why the transitivity of causation has a strong appeal, and why there are nevertheless convincing counterexamples to accepting it. We do so by an appeal to the principle that causation is *asymmetrical*: an event is a cause only if its absence would not have been a cause. Accepting this principle leads the way to an analysis of causation as a transitive relation compromised by asymmetry. This analysis provides us with a sufficient condition for causation to be transitive, a sufficient condition for dependence to be necessary, and several sufficient and necessary conditions for dependence to be transitive. Finally, we show how this analysis leads naturally to several conditions a definition of causation should satisfy. It is then easy to verify that our own definition developed in the previous chapter satisfies these conditions. As a result, our analysis of the transitivity of causation lends additional support to the principles adopted in the previous chapter. The starting point for our analysis consists of a detailed overview of the literature on this topic.

Initially we shall ignore the temporal aspects of causation, which proved essential for cases of Late Preemption. This caveat will not undermine the current discussion, because it stands orthogonal to the issue of transitivity. Therefore after our discussion of transitivity, we integrate our discussion on *production* into our analysis, to show that our definition of actual causation from the previous chapter satisfies both the current analysis and both **Preemption** and **Actual Timing**.

Sections 7.2 offers the relevant background, and presents a survey of the literature on transitivity. This leads us to suggest a first condition any definition of causation should satisfy, in Section 7.3. Section 7.4 presents a sufficient condition for the transitivity of causation. We come back to the asymmetry of causation in Section 7.5 and show how it can be combined with transitivity to form an elegant explanation of all aspects here discussed.

7.2 Literature Survey

In Section 6.2 we presented the sufficiency of counterfactual dependence as the starting point for all approaches under consideration: (Hall, 2004, 2007; Halpern, 2015a; Halpern & Pearl, 2005a; Hitchcock, 2001; Lewis, 1973; Weslake, 2015; Woodward, 2003). We also presented three basic types of examples which are commonly used to defend the claim that dependence is not necessary, namely *Early Preemption*, *Late Preemption*, and *Symmetric Overdetermination*. Of course there exist many more counterexamples to the necessity of dependence, but in essence they can all be reduced to these paradigmatic cases, or combinations thereof. To illustrate, we present another case of *Early Preemption* from Hitchcock (2001)[p. 276]:

Example 17 (Backup)**.** *An assassin-in-training is on his first mission. Trainee is an excellent shot: if he shoots his gun, the bullet will fell Victim. Supervisor is also present, in case Trainee has a last minute loss of nerve (a common affliction among student assassins) and fails to pull the trigger. If Trainee does not shoot, Supervisor will shoot Victim herself. In fact, Trainee performs admirably, firing his gun and killing Victim.*

The following is the standard model used in the literature for this story, where the context is such that $Trainee$ is **true**.

$$Victim := Trainee \vee Supervisor.$$

$$Supervisor := \neg Trainee.$$

Intuitively it is clear that $Trainee$ is a cause of $Victim$, yet using this model we see that $Victim$ is not dependent on $Trainee$. The starting point for any definition of causation in the counterfactual tradition is to provide a way of handling cases of *Early Preemption*. Lewis (1973) does so by invoking another appealing principle of causation: that it is a transitive relation.

Principle 5 (**Transitivity**)**.** *If C is a cause of D and D is a cause of E w.r.t. $(M,\vec{u})$, then C is a cause of E w.r.t. $(M,\vec{u})$.*

Lewis (1973) takes **Dependence** and **Transitivity** at face value, and defines causation as the transitive closure of dependence. This definition is able to handle cases of Early Preemption by focussing on an intermediate event in between Trainee's shot and Victim getting hit, for example the event of the bullet flying through the air. By adding a variable to the model representing this event, say $Bullet$, Lewis gets the desired result: $Victim$ is dependent on $Bullet$, $Bullet$ is dependent on $Trainee$, and thus by **Transitivity**, $Trainee$ causes $Victim$.
Elegant as it may be, McDermott (1995) demonstrated that there are two major problems with this definition: it is neither a necessary condition for causation, nor a sufficient one. For the former: cases of *Late Preemption* and *Symmetric Overdetermination* do not contain a chain of dependencies, and still intuitively exhibit

causation. For the latter: there are intuitively convincing counterexamples to the transitivity of causation. The first problem is generally taken to be a decisive blow to Lewis' definition. The second problem, however, has more general repercussions: giving up **Transitivity** is not taken lightly. To understand why this is the case, we give an overview of the literature on this problem.

7.2.1 Counterexamples to Transitivity

Many authors have tackled the issue of transitivity, and their analysis always contains the following two properties: the transitivity of causation sounds intuitively appealing, but unfortunately there are convincing counterexamples (Hall, 2000, 2004; Halpern, 2015b; Halpern & Pearl, 2005a; Hitchcock, 2001; McDermott, 1995; Paul & Hall, 2013; Sartorio, 2005). Halpern (2015b) is the most recent to take up this view, summarising the importance of transitivity in the counterfactual tradition as follows [p. 2]:

> Paul and Hall (2013)[p. 215] suggest that "preserving transitivity is a basic desideratum for an adequate analysis of causation". Hall (2000) is even more insistent, saying "That causation is, necessarily, a transitive relation on events seems to many a bedrock datum, one of the few indisputable a priori insights we have into the workings of the concept." Lewis (1986, 2000) imposes transitivity in his influential definition of causality, by taking causality to be the transitive closure ("ancestral", in his terminology) of a one-step causal dependence relation.

Although Halpern (2015b) agrees that transitivity should be preserved as much as possible, he acknowledges that there are convincing counterexamples, as do all of the other authors mentioned. To illustrate, we present a few of them here.
The first is by Hitchcock (2001), but Hall (2000) gives an almost identical example.

Example 18 (Boulder)**.** *A boulder is dislodged, and begins rolling ominously toward Hiker. Before it reaches him, Hiker sees the boulder and ducks. The boulder sails harmlessly over his head with nary a centimeter to spare. Hiker survives his ordeal.*

The following is an appropriate model for this story, where the context is such that *Boulder* is **true**.

$$Dies := Boulder \wedge \neg Duck.$$

$$Duck := Boulder.$$

We see that Hiker surviving ($\neg Dies$) is dependent on *Duck*, and *Duck* in turn is dependent on *Boulder*. Hence by **Dependence**, *Boulder* causes *Duck* and *Duck* causes $\neg Dies$. However it would be absurd to conclude from this that the boulder coming down is a cause of hiker's survival. Thus this example presents a violation of **Transitivity**.

The next example is originally due to McDermott (1995)[p. 531], but is also discussed by others (Hall, 2000; Halpern, 2015b; Hitchcock, 2001).

Example 19 (Dog Bite)**.** *Terrorist, who is right-handed, must push a detonator button at noon to set off a bomb. Shortly before noon, he is bitten by a dog on his right hand. Unable to use his right hand, he pushes the detonator with his left hand at noon. The bomb duly explodes.*

We model this as follows, where the context is such that $DogBite$ is **true**.

$$Bomb := LH \vee RH.$$

$$LH := DogBite.$$

$$RH := \neg DogBite.$$

Just as with Boulder, it would be absurd to consider the dog bite to be a cause of the explosion, as implied by **Dependence** and **Transitivity**.
We already presented the *Switch* example in Section 6.5, which is structurally identical to the previous one.

Example 20 (Switch)**.** *An engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the left-hand track, instead of the right. Since the tracks reconverge up ahead, the train arrives at its destination all the same.*

Intuitively, flipping the switch is not a cause of the train's arrival, again going against the combined claims of **Dependence** and **Transitivity**.
Many more counterexamples are given in the literature, but their structures are very similar to the examples here presented. Given the existence of these intuitively convincing counterexamples, all of the authors mentioned agree that **Transitivity** should be abandoned.[1]
Although this means we are sacrificing an intuitive property of causation, we should be careful not to sacrifice too much: even if some cases provide convincing counterexamples to transitivity, there is no reason to abandon it altogether. Again we take our cue from Halpern (2015b)[p. 2]:

> In light of the examples, should we just give up on these intuitions? Paul and Hall (2013) suggest that "What's needed is a more developed story, according to which the inference from "C causes D" and "D causes E" to "C causes E" is safe provided such-and-such conditions obtain – where these conditions can typically be assumed to obtain, except perhaps in odd cases." The goal of this paper is to provide sufficient conditions for causality to be transitive.

[1] Originally Hall did try to hold on to **Transitivity**, by sacrificing **Dependence**. Later, he rejected this view Hall (2000, 2007).

Halpern (2015b) only discusses such conditions in case of dependence. By contrast, we provide several necessary and sufficient conditions for dependence to be transitive, and derive from this a sufficient condition for causation to be transitive *in general*.

7.3 The (In)transitivity of Dependence

By **Dependence** we know already that whenever dependence is transitive, causation will be transitive as well. In all of the papers mentioned, it holds for all of the counterexamples there discussed, that they have one essential thing in common: they are also counterexamples to the transitivity of dependence.[2] This leads to the suggestion that likewise, whenever dependence violates transitivity, so does causation. Taken together this amounts to the following Condition:

Condition 1. *If E depends on D and D depends on C w.r.t. $(M,\vec{u})$, then it holds that: C causes E w.r.t. $(M,\vec{u})$ iff E depends on C w.r.t. $(M,\vec{u})$.*

Any definition which satisfies this condition has the desirable property that it violates transitivity in all of the counterexamples discussed in the literature, while also respecting transitivity in ordinary cases where dependence does so as well.
Recall that dependence is not necessary for causation in general, due to problem cases exhibiting *Early Preemption*, *Late Preemption*, or *Symmetric Overdetermination*. The above condition states that in case we have *a chain of dependencies*, dependence does become a necessary condition. To bring these two observations in agreement requires showing that those problem cases do not occur in case there is a chain of dependencies from C to D to E, but no dependence of E on C.
Regarding *Late Preemption* and *Symmetric Overdetermination*, we shall be brief: there is no example in the literature we know of that is considered a case of either of those, and for which Condition 1 is violated. To illustrate, we come back to our classic example of *Symmetric Overdetermination*:

Example 21. *[Symmetric Overdetermination] Suzy and Billy both throw a rock at a bottle. Both rocks hit the bottle simultaneously, upon which it shatters. Either rock by itself would have sufficed to shatter the bottle.*

We modelled this story using the single equation $BS := ST \vee BT$, and the context such that both ST and BT are **true**. Intuitively both ST and BT are causes of BS, yet BS is not dependent on either of them. It is clear that in this example the failure of dependence has nothing to do with issues of transitivity. Rather, the problem is that there are two completely independent processes which suffice to bring about E, and both of them actually occur, overdetermining E. Adding more detail by inserting variables in between ST and BS, so that there is a chain of causes leading from ST to

[2] This is in line with Hitchcock (2001)[p. 276], who defines *ordinary* cases of causation as those where the transitivity of dependence is respected.

BS, does nothing to change the fact that there will never be a chain of dependencies from *ST* to *BS*.

7.3.1 Early Preemption

Cases of *Early Preemption* provide the other reason why dependence is not necessary for causation. Therefore we need to show that such cases do not occur when there is a chain of dependencies, and no dependence from the end of the chain on its start, so that Condition 1 can be accepted.
There is one basic causal setting that is considered by most to be the prototypical case of Early Preemption: it is the setting used for *Backup* introduced in Section 7.2. The model was the following, where the context is such that C holds:

$$E := C \vee F.$$

$$F := \neg C.$$

Since there is no chain of dependencies, Condition 1 does not apply and there is no problem.
Recall from our discussion of *Early Preemption* from Section 6.6, that we noted the similarity between models for *Switch* and models for *Early Preemption*. This is confirmed here by the observation that the above model is quite similar to that used for *Dog Bite* and *Switch*, two of the counterexamples to the transitivity of causation we discussed earlier. There, the model is as follows:

$$E := D \vee F.$$

$$D := C.$$

$$F := \neg C.$$

As we mentioned in Section 7.2, the account of Lewis exploits this similarity to deal with Early Preemption. Since our aim is to avoid the conclusion that examples like *Dog Bite* and *Switch* are causes, we need to show that there is an alternative way to handle cases of Early Preemption, like *Backup*. We briefly go over several strategies to do so. Note though that at present our goal is not to find the best way to deal with Early Preemption as such, but simply to safeguard Condition 1 while leaving room for a proper analysis of Early Preemption. Concretely, we want to make sure that such an analysis need not conflict with modelling Switches in the manner above. (By calling an example a Switch, we simply mean that it is an example where intuitively there is no causation, as opposed to Early Preemption.)
Our preferred strategy to handle Early Preemption is to use a non-deterministic model, as we showed in Section 6.6. For example, when considering the *Backup* example, we make the plausible assumption that even Supervisors are not always accurate, or may

also have a loss of nerve. The following is an appropriate non-deterministic model for the *Backup* example:

$$Victim := Trainee \vee (Supervisor \wedge Accurate).$$

$$Supervisor := \neg Trainee \wedge \neg Nerves.$$

Here we have added variables to represent the accuracy of Supervisor's shot, and the possibility that he has a loss of nerve. The actual story only gives us the partial context such that $Trainee$ holds, leaving unspecified the values of $Accurate$ and $Nerves$. Applying Definition **??**, we get that $Victim$ is dependent on $Trainee$.

Using a non-deterministic model allows for a counterfactual story in which Victim is not killed in case Trainee does not shoot, implying that $Victim$ is dependent on $Trainee$. (If desired, different causes can be ranked by distinguishing between different levels of dependence. For example, one could add a probability distribution over the exogenous variables.) Whether the model contains an intermediate variable in between C and E or not is immaterial to this strategy.[3] (As an illustration, note that the model for the Early Preemption example from Section 6.6 does contain an intermediate variable, contrary to the model for *Backup*.) Cases of Early Preemption are distinguished from Switches on a one-by-one basis: is the actual story such that the relevant counterfactual story, in which we have both $\neg C$ and $\neg E$, should be part of the model, i.e., is there any reason to doubt that the backup process will function properly? If yes, then there is dependence and we consider it a case of Early Preemption. If no, then the backup process – Supervisor shooting, the functioning of the right hand track, Terrorist using his right hand – is taken to be reliable and it is a Switch. Obviously this distinction involves a certain amount of subjectivity, but given the divergence of intuitions between people regarding the same story we take this to be a benefit of our approach.

Hall (2007) uses a strategy that is similar to ours, but not quite the same. As we saw in Section 5.2.2, he models both Early Preemption and Switches using an intermediate variable, but that is merely a consequence of his use of neuron diagrams. As with our strategy, he agrees that the distinction between the two cases comes down to whether or not the backup process can fail. The difference is that on his view of Early Preemption, even if we somehow have evidence that in the actual story the backup process was reliable and *would not have failed*, (A is in its deviant state in the diagrams in Section 5.2.2), we may still consider the counterfactual story in which it does fail. But, as we showed in Chapter 5, in this view it becomes quite hard – if not impossible – to express the difference between Early Preemption and Switch. For every backup process there is some relevant property on which its reliability depends: Supervisor being accurate or not losing his nerves, Terrorist's ability to use his right hand, the right hand track not being broken, etc. All it takes on his account to change a Switch into a case of Early Preemption is to add a variable representing this property. Then, even if we have

[3] Menzies (2004) also argues that the intermediate variable does not matter, and likewise claims that C is not a cause in either of the two deterministic models.

evidence that the relevant property is present, we may still consider the counterfactual story in which it is not. Because of this undesirable consequence, we do not find this strategy convincing.

Hitchcock (2001) and Halpern and Pearl (2005a) offer a third strategy: although they agree that there is no causation in Switching stories such as *Dog Bite* and *Switch*, they argue that the second model is not appropriate for them. They are forced to take up this position, because their solution to get C to come out as a cause in the first model applies just as well to the second one. In response to the common practice of using the second model for Switches, they argue case by case why on closer inspection that model is not appropriate for a particular story, or why that story should indeed be considered a case of Early Preemption. Let us examine both replies.

Hitchcock (2001) argues against using the model for *Dog Bite*. However, that argument only applies to accounts that make use of so-called "ENF counterfactuals", which are a particular form of interventions on a structural model that we will not go into.[4] Halpern and Hitchcock (2010)[p. 16] argue against using this model for *Switch*, on the basis that the variables LT and RT are logically related: "the train cannot be on both tracks at once". First of all, we disagree that the relation between these variables is logical: it is a matter of physics, not logic, that a train can only occupy a single track at any given moment. Second of all, this argument does not apply to *Dog Bite*, as one can push a detonator using two hands. In light of this, and in absence of a general argument as to why such models should never be used to model Switches, this reply is not convincing.

As mentioned in Section 6.6.1, Halpern and Pearl (2005a)[p. 27] claim that the second model can be appropriate for the *Switch* example, but only if we consider the possibility that the right hand track will fail as relevant. Pearl (2000) makes the same claim regarding a Switch made up of two lamps. The motivation behind this strategy and ours is the same: if we take the failure of the backup process to be a relevant possibility, then we should consider the counterfactual story in which it does. The difference is that we do not seek recourse in structural contingencies (such as ENF counterfactuals) to represent such counterfactual stories, but use a partial context and generalise dependence to include non-deterministic models. It is important to point out that this agreement is limited to the distinction between Switches and Early Preemption, and should not be generalised. Halpern and Pearl (2005a) use structural contingencies to consider vastly different counterfactual stories as well, which have nothing to do with the issue at hand.

Weslake (2015) uses a different strategy altogether. According to him, the models are not similar at all, and only the first should be used to model Early Preemption. He agrees that the second model is a Switch, and therefore C does not cause E.

We have now discussed four of the most important strategies to handling Early Preemption, and presented a number of arguments against adopting the second or the third strategy. What matters for the current investigation, however, is that the first

[4] See Paul and Hall (2013)[ch. 5] for a detailed discussion of the problems these ENF counterfactuals pose to dealing properly with the counterexamples to **Transitivity**.

and fourth strategies are both viable options for handling Early Preemption properly without running into conflict with Condition 1.
To sum up, because of the fact that the counterexamples to **Transitivity** are without exception also counterexamples to the transitivity of dependence, and in light of the lack of opposition from problem cases like Overdetermination and Early Preemption, we claim that Condition 1 should be accepted.

7.4 Transitivity in General

7.4.1 Contributing

A proper understanding of the intransitivity of causation requires looking further than dependence. In the previous chapter we saw that Dependence stands at one end of a spectrum, as a strong but intransitive relation that is sufficient for causation. At the other end there is the concept of *contributing*, which is a weak and transitive relation that is necessary for causation.
The following is an interesting connection between dependence and contributing, that will prove useful for interpreting subsequent results.

Theorem 10. *E depends on C w.r.t. $(M,\vec{u})$ iff C contributes to E w.r.t. $(M,\vec{u})$ and $\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)},\vec{u})$.*

Proof. The implication from right to left is trivial, and the implication from left to right is a direct consequence of Theorem 2 in Chapter 6.

□

7.4.2 A Sufficient Condition for Transitivity

Condition 1 states that causation and dependence are equally transitive in case we have a chain of dependencies.
The next step is to look at transitivity in case there is a chain of causes simpliciter, but not necessarily a chain of dependencies. More specifically, we want to find a good sufficient condition for the transitivity of causation *in general*. Since Condition 1 suffices to respect all counterexamples to **Transitivity** from the literature, a naive suggestion would be to simply demand that causation is always transitive when there is no chain of dependencies. To understand why this would not work, we show how the counterexamples can easily be modified so that there no longer is a chain of dependencies, yet intuition would still find that causation is intransitive. All we need to do is add a little *Symmetric Overdetermination* into the mix.

Example 22 (Dog Bite with Backup)**.** *Imagine the story of the Terrorist from* Dog Bite*, but with a little twist: there are two detonators that can be pushed, either of which will set off the bomb. To make sure nothing goes wrong, Backup pushes the other detonator at the same moment as Terrorist does.*

We can re-use our old model, except that we add Backup's action.

$$Bomb := LH \vee RH \vee Backup$$

$$LH := DogBite.$$

$$RH := \neg DogBite.$$

Just as in the original example, **Dependence** implies that *DogBite* is a cause of *LH*. However, *Bomb* is now no longer dependent on *LH*, so there is no chain of dependencies from *DogBite* to *Bomb* and Condition 1 does not apply. Because *Bomb* is symmetrically overdetermined by both *LH* and *Backup*, we also have that *LH* should still be a cause of *Bomb*. Nevertheless *DogBite* should still not be considered a cause of *Bomb*.
The lesson learned is that the focus should not be on the presence of a chain of dependencies as such, but rather on the conditions that decide whether or not dependence is transitive. Therefore we now present three different characterisations of the transitivity of dependence.

Theorem 11. *If E depends on D and D depends on C w.r.t.* $(M,\vec{u})$*, then the following statements are all equivalent:*

1. *E depends on C w.r.t.* $(M,\vec{u})$.
2. $\neg E$ *depends on* $\neg C$ *w.r.t.* $(M_{do(\neg C)},\vec{u})$.
3. $\neg C$ *contributes to* $\neg E$ *w.r.t.* $(M_{do(\neg C)},\vec{u})$.
4. $\neg C$ *does not contribute to E w.r.t.* $(M_{do(\neg C)},\vec{u})$.

Proof. Assume E depends on D and D depends on C w.r.t. $(M,\vec{u})$. First, note that by Theorem 10, this implies that C contributes to D and D contributes to E w.r.t. $(M,\vec{u})$. Since contributing is transitive by construction, this implies that C contributes to E w.r.t. $(M,\vec{u})$.

$\boxed{1 \Leftrightarrow 2}$

We start with assuming that E depends on C w.r.t. $(M,\vec{u})$. It follows directly from the definitions that this is equivalent to: $\neg E$ depends on $\neg C$ w.r.t. $(M_{do(\neg C)},\vec{u})$.

$\boxed{2 \Leftrightarrow 3}$

Now assume we know that $\neg E$ depends on $\neg C$ w.r.t. $(M_{do(\neg C)},\vec{u})$. Given that we already know that C contributes to E w.r.t. $(M,\vec{u})$, by Theorem 10 we see that this is equivalent to $\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)},\vec{u})$.

$\boxed{3 \Leftrightarrow 4}$

Lastly, assume that $\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)},\vec{u})$. It follows directly that $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)},\vec{u})$. Remains the reverse implication. It suffices to show that if $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)},\vec{u})$, then $(M_{do(\neg C)},\vec{u}) \models \neg E$. We proceed by a reductio: assume that $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)},\vec{u})$ and $(M_{do(\neg C)},\vec{u}) \models E$.

D depends on C w.r.t. $(M,\vec{u})$, and thus $(M_{do(\neg C)},\vec{u}) \models \neg D$. Together with the fact that $(M_{do(\neg C)},\vec{u}) \models E$, this implies that $(M_{do(\neg C,\neg D)},\vec{u}) \models E$. Also, since E depends on D w.r.t. $(M,\vec{u})$, we have $(M_{do(\neg D)},\vec{u}) \models \neg E$. Therefore $\neg E$ depends on C w.r.t. $(M_{do(\neg D)},\vec{u})$. By Theorem 1, this implies that $\neg C$ contributes to E w.r.t. $(M_{do(\neg C,\neg D)},\vec{u})$, and thus also $\neg C$ contributes to E w.r.t. $(M_{do(\neg C)},\vec{u})$, which concludes the proof. □

Note that in general, i.e., without the restriction to chains of dependencies, the statements in this theorem are not all equivalent. Rather we have that $1 \Leftrightarrow 2$, and $2 \Rightarrow 3 \Rightarrow 4$, but also $4 \not\Rightarrow 3 \not\Rightarrow 2$.
This theorem shows that to satisfy Condition 1, it suffices to take any of the three last conditions as a sufficient condition for the transitivity of causation. Since **Transitivity** is intuitively appealing, we want to restrict **Transitivity** as little as possible. Given that the last condition from Theorem 11 is clearly weaker than the other three (in general), this naturally leads to the following condition:

Condition 2. *[Sufficient Condition for Transitivity] If C causes D and D causes E w.r.t. $(M,\vec{u})$, then the following holds:*
If $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)},\vec{u})$ then C causes E w.r.t. $(M,\vec{u})$.

In light of Theorem 11 and **Dependence**, we can interpret Condition 2 informally as stating that a definition of causation ought to be "at least as transitive as dependence", i.e., its sufficiency condition for transitivity should be at least as weak as that for dependence. (Note that this statement can be endorsed without having to accept Condition 1.) Before we follow through on this lead, we present an example to show that violations of Condition 2 lead to counterintuitive results.

7.4.3 Counterexample

Since neither of the definitions from Halpern and Pearl (2005a), Woodward (2003), and Weslake (2015) satisfy Condition 2, we will use them as illustrations.
Consider the following model, with a context such that both A and C hold.

$$E := D \vee A.$$

$$D := C \wedge A.$$

Since D is dependent on C in this setting, by **Dependence** C is a cause of D. We leave it to the reader to verify that according to all three definitions mentioned above, D is a cause of E, even though C is not a cause of E, in violation of **Transitivity**. $\neg C$ does not contribute to E, and thus this example violates Condition 2 as well.
The following story confirms that this is a counterintuitive result.

Example 23 (Assassin)**.** *Assassin adds Cyanide to Victim's coffee, which is certain to kill a person. Backup adds Milk to the coffee, which reacts with the Cyanide to form Arsenic, another lethal substance. Victim drinks his coffee, and dies.*

Our counterexample provides an appropriate model for this story.

$$Dies := Arsenic \vee Cyanide.$$

$$Arsenic := Milk \wedge Cyanide.$$

Obviously Assassin adding the Cyanide is a cause of Victim's death, since *Dies* is dependent on *Cyanide*. If we focus on the first equation, then the situation is identical to *Symmetric Overdetermination*. Given that all three definitions mentioned judge both overdetermining events to be causes of the effect, the same applies here: all three of them also consider *Arsenic* a cause of *Dies*. Further, since *Arsenic* is dependent on *Milk*, Backup adding the Milk was a cause of the Arsenic being in the coffee. However, the above definitions reach the counterintuitive conclusion that despite all this, Backup adding the Milk is not a cause of Victim's death.

7.5 Transitivity and Asymmetry

7.5.1 Asymmetry

In Section 6.5 we presented the asymmetry of causation as an appealing explanation for why *Switch* should not be considered a cause of *Dest* in *Switch*. The same applies to the second model we considered in Section 7.3.1, namely there is a *remarkable symmetry* between the actual story and the counterfactual story that we get when intervening on C: in both cases there is a chain of counterfactual dependence from the candidate cause (C and $\neg C$, respectively) to the effect E. As **Asymmetry** and **Transitivity** focus on entirely different properties of causation, it is no surprise that they conflict with each other:

Theorem 12. Dependence, **Transitivity**, *and* **Asymmetry** *are mutually inconsistent.*

Proof. We have a look again at the Switch example. In the story such that *Switch* holds, by **Dependence** *Switch* is a cause of *LT* and *LT* is a cause of *Dest*. By **Transitivity**, this makes *Switch* a cause of *Dest*. But if we look at the story $do(\neg Switch)$, then we can apply the same reasoning to get that $\neg Switch$ is a cause of *Dest*. This is in violation of **Asymmetry**. □

Theorem 12 teaches us that accepting **Asymmetry** provides an explanation for the fact that there are violations of **Transitivity**.[5] In fact, picking up our earlier discussion, **Asymmetry** together with **Contributing** helps to make sense of Condition 2, which we can rephrase informally as:

> If C causes D and D causes E w.r.t. $(M, \vec{u})$, then the following holds:
>
> **Transitivity** should be respected unless this would violate **Asymmetry**.

[5] Sartorio (2005) also uses Switches to argue that violations of **Transitivity** are due to **Asymmetry**.

There are now enough elements on the table to construct a coherent genesis of causation that explains its limited transitivity.

7.5.2 Putting it all Together

We started our analysis by noting the strong connection between dependence and causation. Specifically, by **Dependence** and **Contributing**, we know that causation lies somewhere in between dependence and contributing. Further, in the overwhelming majority of cases, all three of these concepts behave as a transitive relation. So as a first approximation, we assume causation to be some relation, say $Trans(X,Y)$, which satisfies the following condition:

Condition 3. *1. $Trans(X,Y)$ is transitive.*

2. If $Trans(C,E)$ then C contributes to E w.r.t. $(M,\vec{u})$.

3. If E depends on C then $Trans(C,E)$ w.r.t. $(M,\vec{u})$.

The following generalisation of Theorem 10 offers a useful connection between dependence and such a $Trans(X,Y)$ relation.

Theorem 13. *If $Trans(X,Y)$ satisfies Condition 3, then:*
E depends on C w.r.t. $(M,\vec{u})$ iff $Trans(C,E)$ w.r.t. $(M,\vec{u})$ and $Trans(\neg C,\neg E)$ w.r.t. $(M_{do(\neg C)},\vec{u})$.

Proof. We start with the implication from left to right. So assume E depends on C w.r.t. $(M,\vec{u})$, which is equivalent to $\neg E$ depends on $\neg C$ w.r.t. $(M_{do(\neg C)},\vec{u})$. Hence by applying 3 to both statements, we get the desired result.
Remains the implication from right to left. So assume $Trans(C,E)$ w.r.t. $(M,\vec{u})$ and $Trans(\neg C,\neg E)$ w.r.t. $(M_{do(\neg C)},\vec{u})$. By 2, this implies that C contributes to E w.r.t. $(M,\vec{u})$ and $\neg C$ contributes to $\neg E$ w.r.t. $(M_{do(\neg C)},\vec{u})$. Applying Theorem 10 gives the result. □

The following is a direct consequence of this theorem, which in analogy with **Asymmetry** we may call **Anti-Symmetry**.

Corollary 2 (Anti-Symmetry)**.** *If E depends on C w.r.t. $(M,\vec{u})$, then $\neg E$ depends on $\neg C$ w.r.t. $(M_{do(\neg C)},\vec{u})$.*

Informally, this result tells us that dependence is built up out of any relation $Trans(X,Y)$ that satisfies Condition 3, in conjunction with the constraint that it should be Anti-Symmetrical.
Since for causation we only require **Asymmetry**, the solution is straightforward: causation is built up out of some relation $Trans(X,Y)$ that satisfies Condition 3, in conjunction with **Asymmetry**. Putting all of this together, we get the following tentative characterisation of a good definition of causation:

Condition 4. *There exists a relation $Trans(X,Y)$ such that each of the following holds:*

1. *$Trans(X,Y)$ is transitive.*
2. *C causes E w.r.t. $(M,\vec{u})$ iff $Trans(C,E)$ w.r.t. $(M,\vec{u})$ and $\neg Trans(\neg C,E)$ w.r.t. $(M_{do(\neg C)},\vec{u})$.*
3. *If $Trans(C,E)$ then C contributes to E w.r.t. $(M,\vec{u})$.*
4. *If E depends on C then $Trans(C,E)$ w.r.t. $(M,\vec{u})$.*

Any definition of causation satisfying the first and second part of Condition 4 is a compromise between **Transitivity** and **Asymmetry**: **Transitivity** is sacrificed only to the extent that is required to satisfy **Asymmetry**. Add to this the other two constraints, and we get a definition that has all the properties we have argued for.

Theorem 14. *Any definition of causation satisfying Condition 4 satisfies* **Dependence**, **Asymmetry**, *and* **Contributing**, *and Conditions 1 and 2.*

Proof. **Dependence**: Assume E depends on C w.r.t. $(M,\vec{u})$. By 2, we need to show that $Trans(C,E)$ w.r.t. $(M,\vec{u})$ and $\neg Trans(\neg C,E)$ w.r.t. $(M_{do(\neg C)},\vec{u})$. The former is a direct consequence of 4, so remains the latter.
Since E does not hold in $(M_{do(\neg C)},\vec{u})$, we get that $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)},\vec{u})$. By 3, this implies that $\neg Trans(\neg C,E)$ w.r.t. $(M_{do(\neg C)},\vec{u})$.
Asymmetry and **Contributing** follow immediately from 2 and 3.
Condition 1: Assume E depends on D and D depends on C w.r.t. $(M,\vec{u})$. By 1 and 4 this implies that $Trans(C,E)$ w.r.t. $(M,\vec{u})$. The implication from right to left in the equivalence from Condition 1 follows from **Dependence**. So we need to prove that $\neg Trans(\neg C,E)$ w.r.t. $(M_{do(\neg C)},\vec{u})$ implies that E depends on C w.r.t. $(M,\vec{u})$.
We proceed by a reductio: assume that $\neg Trans(\neg C,E)$ w.r.t. $(M_{do(\neg C)},\vec{u})$ and $(M_{do(\neg C)},\vec{u}) \models E$.
D depends on C w.r.t. $(M,\vec{u})$, and thus $(M_{do(\neg C)},\vec{u}) \models \neg D$. Together with the fact that $(M_{do(\neg C)},\vec{u}) \models E$, this implies that $(M_{do(\neg C,\neg D)},\vec{u}) \models E$. Also, since E depends on D w.r.t. $(M,\vec{u})$, we have $(M_{do(\neg D)},\vec{u}) \models \neg E$. Therefore $\neg E$ depends on C w.r.t. $(M_{do(\neg D)},\vec{u})$. By Theorem 13, this implies that $Trans(\neg C,E)$ w.r.t. $(M_{do(\neg C,\neg D)},\vec{u})$, and thus also $Trans(\neg C,E)$ w.r.t. $(M_{do(\neg C)},\vec{u})$, which concludes the proof.
Condition 2: Assume C causes D and D causes E w.r.t. $(M,\vec{u})$, and $\neg C$ does not contribute to E w.r.t. $(M_{do(\neg C)},\vec{u})$. By 1, we get that $Trans(C,E)$ w.r.t. $(M,\vec{u})$. By 3, we also get that $\neg Trans(\neg C,E)$ w.r.t. $(M_{do(\neg C)},\vec{u})$, and thus C causes E w.r.t. $(M,\vec{u})$. □

Since the weakest possible choice for *Trans* is to take *contributing to*, we state here the most straightforward definition of causation which meets all the demands of Condition 4.

Definition 33 (Actual Causation)**.** *Given* $(M,\vec{u}) \models C \wedge E$*, we define C to be an* actual cause *of E w.r.t.* $(M,\vec{u})$ *if C contributes to E w.r.t.* $(M,\vec{u})$ *and* $\neg C$ *does not contribute to E w.r.t.* $(M_{do(\neg C)},\vec{u})$.

We leave it to the reader to go over all of the examples discussed, to see that this definition gives the desired result. (Again, one should bear in mind our use of non-deterministic models. One counterfactual story in which $\neg C$ does not contribute to E suffices for Definition 33 to be satisfied.)

7.5.3 Coming back to *Production*

As noted, in the current analysis we have ignored the temporal aspects at play in causation. Concretely, we have ignored the crucial distinction between the concepts of *contributing* and *producing* from the previous chapter, that we expressed by extending structural models with a *timing*. Using these notions allowed us to formulate **Preemption**, stating that causes must come before their effects. Adding this aspect to the current analysis is straightforward: in Condition 4, extend the causal settings with a timing τ, and replace the third part with "If $Trans(C,E)$ then C produces E w.r.t. $(M,\vec{u},\tau)$". This ensures that **Preemption** is satisfied as well.
Now the weakest possible choice for *Trans* is to take *producing*. As a result, the most straightforward definition of causation that meets all the demands of the updated Condition 4 is precisely the definition we proposed in Chapter 6:

Definition 34. *Given* $(M,\vec{u},\tau) \models C \wedge E$*, we define C to be an* actual cause *of E w.r.t.* $(M,\vec{u},\tau)$ *if C produces E w.r.t.* $(M,\vec{u},\tau)$ *and* $\neg C$ *does not produce E w.r.t.* $(M_{do(\neg C)},\vec{u},\tau_{do(\neg C)})$.

7.6 Conclusion

Starting from the observation that despite the intuitive appeal of the transitivity of causation there are many convincing counterexamples to accepting it, we have constructed an analysis in order to explain the precise relation between causation and transitivity. By pointing out the connection between violations of the transitivity of dependence, and violations of transitivity in general, we arrived at a characterisation of the transitivity of dependence that suggested a suitable sufficient condition for the transitivity of causation. Adding to this the principle of asymmetry resulted in a detailed genesis of causation, that narrows down the search to a proper definition of causation considerably. Using this we have suggested a definition which meets all the requirements discussed. Finally, we complemented the analysis of transitivity with the temporal properties of causation discussed in the previous chapter, to arrive at the definition of actual causation we suggested in Chapter 6. Thus our analysis of transitivity offers further support for the definition we have developed.

Chapter 8

Conclusion

This work has been an investigation into the concept of actual causation. First, we have developed a flexible, general framework, that can be used to construct various definitions of actual causation, as well as extensions to those definitions that incorporate judgments of normality. Second, we explored actual causation starting from basic principles, and used those as building blocks in constructing a definition of causation. Concretely, the contributions of this work are the following.

In Chapter 2 several formal languages were introduced: the structural equations framework (Pearl, 2000), CP-logic (Causal Probabilistic logic) (Vennekens et al., 2009), and neuron diagrams as used by Hall (2004). We also added the most influential definition of actual causation from Halpern and Pearl (2005a). Lastly, we presented a translation between these three frameworks.

8.1 A General Framework for Defining and Extending Actual causation using CP-logic

In Chapter 3 we used CP-logic to develop a general parametrised definition of actual causation. Although it expresses a probabilistic degree of causation, it continues the counterfactual tradition of causation that started with Lewis (1973). We presented four definitions of causation as instantiations of this general definition, based on definitions by Hall (2004, 2007); Vennekens (2011) and Beckers and Vennekens (2012). This allowed us to compare these definitions directly, leading to the fundamental idea that a definition of causation should be a suitable compromise between *dependence* and *production*.

In Chapter 4 we extended the general definition from the previous chapter, so as to incorporate recent findings from the psychology literature on the context-sensitivity of causal judgments (Hitchcock & Knobe, 2009; Knobe & Fraser, 2008; Moore, 2009).

We based ourselves on a similar extension by Halpern and Hitchcock (2015). First we translated their work from structural equations into CP-logic, and then added several improvements that the quantified setting of CP-logic allows for.
Chapter 5 took a look at three problems facing the two main definitions of causation from Chapter 3. The first problem pertains to the Hall07 definition, and consists in an undesirable sensitivity to details of the model. In his criticism of Hall's definition, Hitchcock (2009) comes to a very similar conclusion. The second problem was pointed out by Hitchcock as well, and applies to both the Hall07 and the BV12 definitions. Finally, the third problem highlights the failure of the BV12 definition to satisfy a fundamental principle regarding causation.

8.2 A Principled Approach to Defining Actual Causation

Taking note of the problems discussed at the end of Part I, in Chapter 6 we started with a clean slate, and proceeded to construct a novel definition of causation using structural equations, extended with a timing. We did so from the bottom up, letting ourselves be guided by paradigmatic examples to discover basic principles which a definition should satisfy. Some of these principles have also been defended (in slightly different forms) by Lewis (1973); Weslake (2015) and Sartorio (2005). Again we arrived at the conclusion that an adequate definition of causation forms a compromise between the concepts of *dependence* and *production*.
Chapter 7 investigated one fundamental principle in particular, that has played a pivotal role in defining causation ever since the work of Lewis (1973): the transitivity of causation. Reluctant to either accept or reject it, most accounts have struggled to give it a proper place (Hall, 2000, 2004; Halpern, 2015b; Halpern & Pearl, 2005a; Hitchcock, 2001; McDermott, 1995; Paul & Hall, 2013; Sartorio, 2005). Using the insights from Chapter 6, we provided a detailed analysis of both the appeal and the problems with transitivity. As a result, we came up with several interesting conditions which a definition of causation should satisfy in order to avoid the pitfalls posed by transitivity. Finally, these conditions lend further support to the definition developed in Chapter 6.

Bibliography

Beckers, S., & Vennekens, J. (2012). Counterfactual dependency and actual causation in cp-logic and structural models: a comparison. In *Proceedings of the sixth stairs* (pp. 35–46). doi: 10.3233/978-1-61499-096-3-35

Beckers, S., & Vennekens, J. (2015a). Combining probabilistic, causal, and normative reasoning in cp-logic. In *12th international symposium on logical formalizations of commonsense reasoning* (pp. 32–38).

Beckers, S., & Vennekens, J. (2015b). Towards a general framework for actual causation using cp-logic. In *Proceedings of the 2nd international workshop on probabilistic logic programming co-located with iclp* (Vol. 1413, pp. 19–38).

Beckers, S., & Vennekens, J. (2016a). A general framework for defining and extending actual causation using cp-logic. *International Journal for Approximate Reasoning*, *77*, 105–126.

Beckers, S., & Vennekens, J. (2016b). A principled approach to defining actual causation. *Synthese*, *forthcoming*.

Collins, J. (2000). Preemptive prevention. *Journal of Philosophy*, *97*(4), 223–234.

Fenton-Glynn, L. (2015). A proposed probabilistic extension of the halpern and pearl definition of 'actual cause'. *The British Journal for the Philosophy of Science*.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787).

Hall, N. (2000). Causation and the price of transitivity. *Journal of Philosophy*, *97*(4), 198-222.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (p. 225-276). The MIT Press.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies*, *132*(1), 109-136.

Hall, N., & Paul, L. A. (2003). Causation and preemption. In P. Clark & K. Hawley (Eds.), *Philosophy of science today.* Oxford University Press.

Halpern, J. (2015a). A modification of the halpern-pearl definition of causality. In *Proceedings of the 24th ijcai* (pp. 3022–3033). AAAI Press.

Halpern, J. (2015b). Sufficient conditions for causality to be transitive. *Philosophy of*

Science.

Halpern, J., & Hitchcock, C. (2010). Actual causation and the art of modeling. In *Causality, probability, and heuristics: A tribute to judea pearl* (p. 383-406). London: College Publications.

Halpern, J., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*, *66*(2), 413–457.

Halpern, J., & Pearl, J. (2005a). Causes and explanations: A structural-model approach. part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–87.

Halpern, J., & Pearl, J. (2005b). Causes and explanations: A structural-model approach. part II: Explanations. *The British Journal for the Philosophy of Science*, *56*(4).

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy*, *98*, 273-299.

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *The Philosophical review*, *116*(4), 495–532.

Hitchcock, C. (2009). Structural equations and causation: six counterexamples. *Philosophical Studies*, *144*, 391-401.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, *106*, 587-612.

Hume, D. (1748). *An enquiry concerning human understanding* (P. Millican, Ed.).

Kahneman, D. T., Daniel; Miller. (1986). Norm theory: comparing reality to its alternatives. *Psychological Review*, *94*(2), 136-153.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology.* MIT Press.

Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*, 113–126.

Lewis, D. (1986). Causation. In *Philosophical papers ii* (p. 159-213). Oxford University Press.

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

McDermott, M. (1995). Redundant causation. *The British Journal for the Philosophy of Science*, *46*(4), 523–544.

Menzies, P. (2004). Causal models, token causation, and processes. *Philosophy of Science*, *71*(5), 820-832.

Moore, M. S. (2009). *Causation and responsibility*. OUP Oxford.

Paul, L., & Hall, N. (2013). *Causation: a user's guide*. Oxford University Press.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Sartorio, C. (2005). Causes as difference-makers. *Philosophical Studies*, *123*, 71–96.

Sato, T. (1995). A statistical learning method for logic programs with distribution semantics. In *Proceedings of the 12th international conference on logic programming* (pp. 715–729).

Shafer, G. (1996). *The art of causal conjecture*. MIT Press. Retrieved from `http://books.google.be/books?id=sY7os7OCykUC`

Vennekens, J. (2011). Actual causation in cp-logic. *Theory and Practice of Logic*

Programming, *11*, 647-662.

Vennekens, J., Denecker, M., & Bruynooghe, M. (2009). CP-logic: A language of probabilistic causal laws and its relation to logic programming. *Theory and Practice of Logic Programming*, *9*, 245-308.

Vennekens, J., Denecker, M., & Bruynooghe, M. (2010). Embracing events in causal modelling: Interventions and counterfactuals in CP-logic. In *Jelia* (pp. 313–325).

Weslake, B. (2015). A partial theory of actual causation. *The British Journal for the Philosophy of Science*, *forthcoming*.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
DTAI
Celestijnenlaan 200A
B-3001 Leuven
Sander.Beckers@cs.kuleuven.be
http://www.dtai.cs.kuleuven.be