# Where causality, conditionals and epistemology meet

*A logical inquiry*

Xie, K.

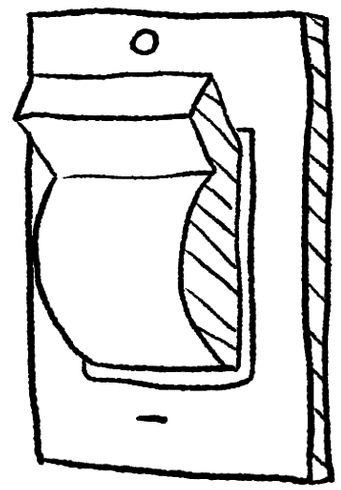
INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Kaibo Xie

Where Causality, Conditionals and Epistemology meet

# Where Causality, Conditionals and Epistemology meet;

## A Logical Inquiry



Kaibo Xie

# Where Causality, Conditionals and Epistemology meet; A Logical Inquiry

**Kaibo Xie**

# Where Causality, Conditionals and Epistemology meet; A Logical Inquiry

ILLC Dissertation Series DS-200X-NN

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

We acknowledge the generous support of a 4-year Chinese Scholarship Council (CSC) scholarship which made this PhD thesis possible and also thank the UvA-Tsinghua Joint Research Center for Logic for their support.



Cover design by Kaibo Xie.
Printed and bound by Ipskamp.

ISBN: 978-94-6421-113-9

# Where Causality, Conditionals and Epistemology meet; A Logical Inquiry

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties
ingestelde
commissie, in het openbaar te verdedigen
26 November 2020

door

Kaibo Xie

geboren te Hangzhou

**Promotiecommisie**

| | | |
|---|---|---|
| Promotor: | prof.dr. S.J.L Smets | Universiteit van Amsterdam |
| Co-promotor: | dr. K. Schulz | Universiteit van Amsterdam |
| | prof. dr. F. Liu | Universiteit van Amsterdam |
| Overige leden: | prof. dr. H.D. Rott | Universität Regensburg |
| | dr. T.F. Icard III | Stanford University |
| | dr. J. Zhang | Lingnan University, Hongkong |
| | prof. dr. F.J.M.M. Veltman | Universiteit van Amsterdam |
| | prof. dr. J.F.A.K. van Benthem | Universiteit van Amsterdam |
| | prof. dr. Y. Venema | Universiteit van Amsterdam |
| | dr. M.D. Aloni | Universiteit van Amsterdam |

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

# Acknowledgments

My first and most important word of thanks goes to my promotor Sonja Smets and main supervisor Katrin Schulz for their invaluable supervision and guidance. They were always correcting and commenting in depth on my drafts. My dissertation benefited greatly from their extensive and inspiring feedback. I learned a lot about how to do fruitful collaborative research by working on joint paper with them. The topics I was working on were closely related to the fields in which Sonja and Katrin are experts. In the regular meetings I always benefited from Sonja's deep insight from the perspective of epistemic logic, and Katrin's profound knowledge of the literature in formal semantics and causality. I really appreciate Sonja and Katrin's kindness and patience during the writing phase of my dissertation. Due to the coronavirus pandemic, I had to work from home for months during the writing stage of the thesis. I cannot imagine how I could have gotten through the tough period without the encouragement from them. In addition, many times they helped me with administrative issues far beyond their responsibility. Their support was indispensable when it came to finishing my PhD.

I want to express my deep gratitude to my external supervisor Fenrong Liu, who has kept giving me precious suggestions and comments since I started studying logic. The first time I met Fenrong was during my undergraduate study. She suggested me to attend EASSLLI 2012 summer school, where Jouko Väänänen, Dag Westerståhl and many other famous researchers taught. As I was an undergraduate student majoring in philosophy at that time, it was difficult for me to follow the lectures, but this became my starting point in logic study. She kept encouraging me to follow cutting-edge topics, and provides me with precious comments on the direction of my research. Her advice and encouragement was crucial in my choice of studying logic and finishing PhD at the ILLC.

I am also grateful to Johan van Benthem. During his visit to China in 2012 for a short-term course, introduced by Fenrong, I was luckily offered an

opportunity to meet with him in person after his course. In the meeting, I was introduced to a variety of interesting topics in the field of logic, of which I was previously unaware. Later, during Johan's teaching in Tsinghua University, I benefited a lot from his lectures. Johan kept providing me with a lot of precious suggestions and feedback on my work during my master and PhD. Many ideas resulting from my discussions with him play an important role in the development of this thesis.

During my studies at ILLC, I benefited greatly from discussions with other faculty members, master students and visiting scholars at ILLC, including Chengwei Shi, Dazhu Li, Fernando R. Velázquez Quesada, Fausto Barbero, Alexandru Baltag, Robert van Rooij, Xiaoxuan Fu, Gianluca Grilletti, Kai Li, Yu Wei, Dean McHugh, Fengkui Ju, Peter van Emde Boas, Bonan Zhao, Frederik Lauridsen, Tim Henke, Anthia Solaki, Ilaria Canavotto, Levin Hornischer, Thom van Gessel, Tom Schoonen, Aybüke Özgün and many others. Discussion was my favorite part of my academic life at the ILLC, in which I not only learnt a lot from the them, but also shared and refined my own ideas. Many comments and inspiring remarks that resulted from those enjoyable discussions are crucial for my thesis. The circle of researchers who contributed to my dissertation indeed goes beyond the ILLC. It is a pity that due to the space limit I cannot list all of them, but I will always remember everybody who contributed directly or indirectly to my thesis.

I am particularly indebted to the co-authors of the papers included in this thesis. I benefited a lot from collaborating with them, in which they shared their invaluable ideas based on the fields in which they are experts.

I owe my special thank to one of my officemates Dean McHugh, who not only inspired me through the discussions in the office, but also kindly helped me in proofreading some of my papers.

I want to express my sincere thanks to the committee members, Hans Rott, Thomas Icard, Jiji Zhang, Frank Veltman, Johan van Benthem, Yde Venema, Maria Aloni for spending their valuable time reading my dissertation.

I am grateful to the ILLC, which provided a warm and friendly environment in which to research. I am particularly grateful to Jenny Batson, Karine Gigengack, Tanja Kassenaar, Debbie Klaassen and Caitlin Boonstra for helping me with various practical issues.

I am grateful to the Tsinghua-UvA Joint Research Center for Logic. By inviting the researchers from the ILLC to Tsinghua University. The joint center provided me with many opportunities to learn from the top research in the world when I was in Tsinghua via joint workshop, regular meetings and lectures. My experience in the joint center during my master study was crucial in making my choice to study at the University of Amsterdam. After I went to the ILLC, the joint center is a bridge for me to collaborate with other researchers in Tsinghua.

I also want to thank Beishui Liao, who supervised my undergraduate thesis,

and Lu Wang, who supervised my master thesis. Their guidance was indispensable in my journey to study logic. They keep providing me with invaluable advice on my research and academic career.

I am also grateful to the CSC (China Scholarship Council) Funding for providing me with financial support during my PhD study in the Netherlands.

Finally, I want to thank all of my family members, who support me in many ways. My parents always encourage me to concentrate on things I am interested in, and give me the freedom to make my own choices. I would not have achieve my PhD degree without their understanding and encouragement.

# Contents

# Chapter 1

# Introduction

## 1.1 Causality, conditionals and epistemic states

This thesis is a philosophical investigation which locates itself at the intersection of studies of causality, epistemology and conditionals. By focusing on a couple of specific issues at this intersection we hope to advance our understanding of causal relations themselves, the way we talk about them in conditionals and our epistemic attitudes toward them. Causality, epistemology and conditionals are three central topics in the philosophical debate. Let us start by saying a bit more about each of them and the way they are related before introducing the main questions the thesis focuses on.

What a conditional sentence expresses has been a central question in the philosophy of language for a long time. The truth condition of a conditional sentence "If A then B" is obviously different from the material implication $A \rightarrow B$. But it turns out to be very difficult to come up with a formal approach that correctly captures these truth conditions. Even though it has been a central objective in the work of many philosopher, still no consensus has been reached about which approach should be chosen.

A maybe even older question in the philosophical debate is the question about the nature of causality. What kind of metaphysical foundation does it have? How do we come to know about causal dependencies? How do we reason with this knowledge? The investigation on causality starts with Aristotle, has been a central topic for philosophers like Kant and Hume and also dominates many current debates in philosophy, cognition and artificial intelligence. In recent years we have seen great progress in formal approaches to causal reasoning and causal learning (Pearl [2002] and Halpern [2000]), but still are a long way from answering the underlying philosophical questions.

Also epistemological questions have a long tradition in the philosophical debate. Next to the traditional questions about the nature of knowledge and belief in more recent times a lot of work has focused on the mechanisms of change

in the epistemic states of agents, based on interaction with the environment and other agents (See for instance Baltag and Smets [2008b], van Benthem [2011] ). Also this research went beyond the boarders of philosophy and has various applications in computer science, artificial intelligence and media studies.

Conditionals, causality of epistemic states are also closely related to each other. First, there seems to be a surprising strong connection between causality and the meaning of a particular group of conditionals: counterfactual conditionals. These are conditionals reasoning about hypothetical situations that are – to the reasoner – known to be false. Some philosophers, most famously Lewis have even argued that the question of causal dependence can be reduced to the truth conditions of certain corresponding counterfactual conditionals.

Second, there appears to be also a very close relation between the meaning of conditionals and belief change. The most dominant approach to the meaning of conditionals follows Ramsey's proposal, according to which we evaluate a conditional by first updating our belief state with the belief in the antecedent and then check if this new belief state supports the belief in the consequent of the conditional. There is a lot of debate in the field on the question whether this is also the right way to approach the meaning of counterfactual conditionals. The causal approaches to the meaning of these sentences that became very popular in recent years follow a different direction. But there are a couple of famous examples in the literature that seem to force the conclusion that also counterfactuals are interpreted based on reasoning about belief change. These examples will play a central role in Chapter 3 of this thesis.

Finally, some of the core questions concerning causality also have a strong epistemological dimension. Think, for instance, of the issue of how we come to know about causal dependencies. Central in this thesis will be the question of how we reason with the causal knowledge or beliefs that we have, and how this then might affect which conditional sentences we are willing to endorse.

In recent years we can see that with the help of formal tools a lot of progress has been made in the study of conditionals, causality and epistemology separately. The development of the causal modelling approach has let to a flood of new developments on the topic of causation. This framework has been also used very successfully for the semantics of conditional sentences (Pearl [2013], Schulz [2011], Kaufmann [2013], Halpern [2016] and many more). Independently, the development of new formal tools like Dynamic Epistemic Logic has let to a lot of progress on epistemological questions. The central innovation of this thesis lies is bringing these new formal tools together to study issues and open problems concerning the interaction between causality, epistemology and conditionals. Because these three topics are intrinsically related, drawing from work on causal modelling, the analysis of conditionals and counterfactuals and the modelling of epistemic attitudes, allows us to establish new connections and provides us with a deeper insight into a number of open issues in the literature. In particular in this thesis we will focus on the following four main

questions:

- Can we build a qualitative formal system that can handle both causal and epistemic reasoning together? Such a system should provide a syntax and semantics that combines both counterfactual and qualitative epistemic operators.
- Can we design a unified account of counterfactuals that can treat both, causal and epistemic examples without having to postulate an ambiguity?
- Can we provide a causal modelling approach to the meaning of counterfactual conditionals that can account for recent observations concerning right-nested counterfactuals?
- Can a causal analysis of conditionals shed light on Aristotle's modal syllogistic?

To answer these questions we will combine philosophical analysis with methods from logic and formal reasoning while using input from experimental results and examples in the literature.

## 1.2 The structure of the dissertation

The thesis is a collection of four research papers. Each of these research papers has either been published or is currently under review. At and has in the exact same form been copied into a chapter in this thesis (while for chapters 2 and 4 we have added the appendix of the paper at the end).

**Chapter 2.** In chapter 2, we will build a bridge between the work on causal models/logics and epistemic models/logics. As mentioned before, in recent years there has been a lot of development on the design of formal models for causal reasoning. Especially the causal models developed by Pearl and Halpern stand out. In a causal model, reality is characterized by a set of causal variables. The causal dependencies between these variables are presented by a set of structural functions connecting the values of the variables. However the model does not represent the epistemic state of an agent. Therefore, the causal language based accompanying the causal model cannot directly express reasoning about causal knowledge. In order to represent epistemic states in causal reasoning, we will extend the notion of a causal model to that of an epistemic causal model. This model will represent both epistemic and causal information. We will develop a corresponding logic for this new notion of model and discuss some of its central properties. We will also address its relationship to other recent extensions of causal models.

- Chapter 2 was written by K. Xie, F. Barbero, K. Schulz, S. Smets and F. R. Velázquez-Quesada and has been submitted to DALI conference 2020 Prague.
- Contribution: K. Xie proposed a formal framework and arguments which were then further developed and extended in collaboration with the co-authors.

**Chapter 3.** In the third chapter, we will show that a combination of causal and epistemic logic allows us to provide a better understanding of the semantics of counterfactual conditionals. As has been shown by various authors, a causal analysis is very successfull in accounting for the meaning of counterfactuals. However, it seems that such an analysis is also limited in certain ways. In particular, there appears to be a group of examples that such an approach cannot account for. These examples have been often claimed to involve an epistemic interpretation of counterfactuals. We will show that extending the causal modeling approach with a representation of the knowledge and beliefs of an agent allows us to provide a unified analysis of counterfactuals, which can deal with these problematic examples.

- Chapter 3 has been written by K. Xie and K. Schulz and has been submitted to *Erkenntnis*.
- Contribution: K. Xie provided the formal framework and developed the argumentation and examples in close cooperation with K. Schulz

**Chapter 4.** Also the fourth chapter focuses on the meaning of counterfactual conditionals. We will discuss a recent challenge brought forward against the standard causal approach to the meaning of counterfactuals. According to this objection, this approach cannot in general account for the interpretation of right-nested counterfactuals, the problem being its strict interventionism. We will report on the results of an empirical study supporting the objection and extend the well-known logic of actual causality with a new operator expressing an alternative notion of intervention that does not suffer from the problem (and thus can account for the critical examples). The core idea of the alternative approach is that the new notion of intervention operates on the evaluation of the variables in a causal model, and not on their functional dependencies.

- Chapter 4 is a reprint of K. Schulz, S. Smets, F. R. Velázquez-Quesada, K. Xie (2019). A logical and empirical study of right-nested counterfactuals. In *Proceedings of Logic, Rationality, and Interaction; 7th International Workshop*. Lecture Notes in Computer Science book series (LNCS, volume 11813), Pages 259-272.

- Contribution: K.Xie initiated the paper with formal framework and arguments which were then further developed and extended in collaboration with the co-authors. The experimental results were brought in by K. Schulz.

**Chapter 5.** The fifth chapter of the thesis aims to investigate a particular application of the causal analysis of conditionals. In this chapter, we will provide a causal analysis of modal syllogism. Modal syllogisms are first discussed in Aristole's Prior Analysis. Aristole claims that some forms of modal syllogisms are valid and some are not, without that it is clear how he arrives at these judgements. For a very long time, the interpretation of modal syllogism has been considered a puzzle: what is the right semantics for modal syllogisms such that it could fit Aristotle's claims for their validity. Many commentators even consider that there is no consistent way to provide a semantics for Aristotle's modal syllogisms. In this chapter we try to solve this problem. The core idea of our approach is to provide a causal semantics for the conditionals involved in modal syllogisms. We argue that this formalization fits better with Aristotle's intuitions concerning the evaluation of modal syllogisms.

- Chapter 5 is a reprint of R. van Rooij, K. Xie, A causal analysis of modal syllogisms, Second Tsinghua Interdisciplinary Workshop on Logic, Language, and Meaning: Monotonicity in Logic and Language.The paper has has been accepted in the proceeding of the TLLM 2020 workshop, to be published in the FoLLI LNCS series after revision.
- Contribution: K. Xie contributed to the formal analysis and part of the writing.

### 1.2.1 Technical preliminaries

In this section we provide a very basic introduction to the main formal frameworks we will employ in the thesis. We can keep this part short, because detailed introductions will be provided in each of the individual chapters.

**Causal models.** The notion of causal models has been developed in Galles and Pearl [1998a], Pearl [2002], Halpern [2000] and Glymour and Spirtes [1988]. The goal of this type of model is to provide a representation of the available information about general causal dependencies. This model then has been used to formalize causal reasoning, but also to serve as foundation for approaches to causal learning.

A causal model is a pair $(\mathcal{S}, \mathcal{F})$. The signature $\mathcal{S}$ consists of a set of causal variables, which represent the entities in the causal scenario to be modelled. These can be of various types, a particular stance as to the objects linked

by causal dependencies is not presumed. A signature is defined as a triple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$ where $\mathcal{U}$ represents the variables that are causally independent of other variables (known as exogenous variables), $\mathcal{V}$ represents the variables that causally depend on some other variables (known as endogenous variables) and $\mathcal{R}$ is a function that indicates the range of possible values of each variable. We can understand a causal model as zooming in on one particular part of the overall network of causal dependencies that governs reality as we see it. For some variables their causes are not part of the particular section of reality we are zooming in on. These become the exogenous variables of the model.

The causal dependencies among the variables are represented by a set of structural functions $\mathcal{F}$, which map each endogenous variable $Y$ to a function $F_Y$ that determines the value of $Y$ given the value of all the other variables. So, they describe the exact nature of the causal laws. They allow us to calculate the exact status (value) of the effect given the status (value) of its cause(s)

**Dynamic epistemic logic.** Dynamic epistemic logic (DEL) is a logical approach for the analysis of epistemic and doxastic attitudes which not only has logic operators that describe epistemic states, but also has dynamic operators to characterize model changes resulting from information change. The earlier work in this direction dates back to Hinttika's epistemic logic (Hintikka [1973] ), and was further developed in a new direction by several researchers, e.g. Plaza [2007], Baltag et al. [1998], van Ditmarsch et al. [2008], Baltag and Smets [2008b], Liu [2007], Van Ditmarsch [2005], van Benthem [2011]). The language of epistemic logic is able to express the epistemic state of an agent by epistemic operators. $K\phi$ stands for the agent knows $\phi$ and $Bel\phi$ stands for the agent believes $\phi$.

According to the setting in Baltag and Smets [2008b] , an epistemic (plausibility) model is a tuple $(\langle W, V, \trianglelefteq\rangle, w)$ where $W$ is a set of possible worlds, $V$ is a valuation function, and $\trianglelefteq$ is a well preorder on $W$[1], $w \in W$ is the actual world. $\trianglelefteq$ is known as the plausibility ordering: $w_1 \trianglelefteq w_2$ means that $w_1$ is at least as plausible as $w_2$ to the agent. If neither $w_1 \trianglelefteq w_2$ nor $w_2 \trianglelefteq w_1$, then $w_1$ and $w_2$ are two epistemically distinguishable worlds for the agent[2].

The semantics is that an agent knows $\phi$ ($K\phi$) if and only if $\phi$ is true in all the epistemically indistinguishable worlds from the actual world. An agent believes $\phi$ ($Bel\phi$) if and only if $\phi$ is true in all the most plausible worlds from the actual world.

Given the model illustrated in figure 1.2.1, let $p$ and $q$ be two propositions and let us assume that in the actual world $\neg p$ and $q$ are the case. Figure 1.2.1, characterizes a situation in which the agent knows that $\neg p$ is the case, while

[1] this is a connected, transitive and well founded relation

[2] Note that in the epistemic models used in chapter 3, we will make the information partition which is induced by the plausibility relation formally explicit in the signature of the models.

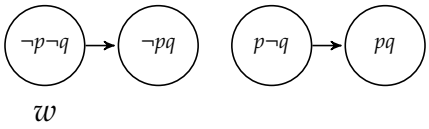



Figure 1.1: An example of an epistemic model

she believes $\neg q$ (we assume that $w$ is the actual world).

For details and explanations on the use of the dynamic operators in DEL as well as applications of the setting to different scenarios, we refer the reader to Baltag and Renne [2016].

# Chapter 2

# Thinking about Causation: a causal language with epistemic operators

## 2.1 Introduction

In recent years a lot of effort has been put in the development of formal models of causal reasoning. A central motivation behind this is the importance of causal reasoning for AI. Making computers take into account causal information is currently one of the central challenges of AI research Pearl and Mackenzie [2019], Bergstein [2020]. There has also been tremendous progress in this direction after the earlier groundbreaking work in Pearl [2000] and Spirtes et al. [1993]. Advanced formal and computational tools have been developed for modelling causal reasoning and learning causal information, with applications in many different scientific areas. In this paper we want to extend this work further. The direction we want to explore is that of developing formal models of the interaction between causal and epistemic reasoning.

Even though the standard logical approach to causal reasoning (Pearl [2000], Halpern [2000, 2016]) can model epistemic uncertainty[1], it does not permit reasoning about the interaction between causal and epistemic reasoning in the object language. Even though recently there have been proposals adding probabilistic expressions to the object language (e.g., Ibeling and Icard [2020]), very little has been done on combining causal and qualitative epistemic reasoning.[2] However, this kind of reasoning occurs frequently in our daily life, especially in connection with counterfactual thinking. Consider, for instance, the following situation.

[1] This can be achieved by adding a probability distribution over the exogenous variables of a causal model. Uncertainty is then restricted to the value of variables. All causal dependencies are deterministic.

[2] See Barbero and Sandu [2019] for an exception, though the epistemic element is not made fully explicit in the language. Section 2.6 discusses the relationship between the referred paper and the current proposal.

**1.** EXAMPLE. *Suppose there is a button in front of Billie, which is connected with a circuit breaker and a sprinkler. If the circuit is closed, then the sprinkler is working if and only if the button is pushed. If the circuit is not closed, the sprinkler won't work, independently of the state of the button. Billie knows the causal laws in this scenario. She can also see the button and the sprinkler, but she has no idea what the state of the circuit breaker is. Suppose furthermore, that at the moment the circuit is closed, the button is not pushed, and the sprinkler, as a result, is not working.*

In such a situation we want to be able to derive that Billie is not sure that if the button had been pushed, the sprinkler would have been working. Thus, we want to be able to make inferences involving epistemic attitudes towards counterfactuals, which, in turn explore causal dependencies. We also want to be able to reason counterfactually about such epistemic attitudes. Considering the same example, we also want to be able to infer that if Billie had pushed the button and saw that that the sprinkler works, then she would have known that the circuit is closed (because of the causal knowledge she has). In order to formalize this type of reasoning, we need a framework that combines causal reasoning with a model of epistemic attitudes.

Given the vast literature on epistemic logic, there is a lot of work that we can build on. This paper makes a start on combining the standard approach to causal reasoning (Pearl [2000], Halpern [2000, 2016]) with tools from Dynamic Epistemic Logic (DEL; Baltag et al. [1998], van Benthem [2011], van Ditmarsch et al. [2008]). The main motivation for this choice is the dynamic character of both systems, even though this aspect will not be explored at depth here. For now we will only consider a very simple extension of the standard system of causal reasoning. But, as we will show, this basic extension already allows us to formalise some interesting concepts and formulate concrete questions for further research.

After introducing the standard approach to causal reasoning in Section 2.2, we will explain in more detail, in Section 2.3, the particular aspects of the interaction between causal and epistemic reasoning we intend to capture here. Section 4.5 will contain the core of the paper: we will extend the standard causal modeling approach in order to deal with with knowledge and external communication. In Section 2.5 we will provide a sound and complete axiomatization of this logic. We conclude the paper discussing the relationship between our proposal and other recent extensions of causal models.

## 2.2 The standard causal modelling approach

What we refer to as the standard logic of causal reasoning was presented on Pearl [1995], extended in Galles and Pearl [1998a], and then further developed in, among others, Halpern [2000], Pearl [2002], Briggs [2012]. This section recall briefly the most important concepts and tools.

The starting point is a formal representation of causal dependencies. This is done in terms of causal models, which represent the causal relationships between a finite set of variables. These variables, as well as their ranges of values, are given by a *signature*. Throughout this text, let $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ be a *finite* signature where

- $\mathcal{U} = \{U_1, \ldots, U_m\}$ is a finite set of *exogenous* variables (those whose value is causally independent from the value of every other variable in the system),
- $\mathcal{V} = \{V_1, \ldots, V_n\}$ is a finite set of *endogenous* variables (those whose value is completely determined by the value of other variables in the system), and
- $\mathcal{R}(X)$ is the non-empty range of the variable $X \in \mathcal{U} \cup \mathcal{V}$.[3]

A causal model is formally defined as follows.

**1.** Definition (Causal model). *A* causal model *is a triple* $\langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$ *where*

- $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ *is the model's signature,*
- $\mathcal{F} = \{f_{V_j} \mid V_j \in \mathcal{V}\}$ *assigns, to each endogenous variable* $V_j$*, a map*

$$f_{V_j} : \mathcal{R}(U_1, \ldots, U_m, V_1, \ldots, V_{j-1}, V_{j+1}, \ldots, V_n) \to \mathcal{R}(V_j).$$

  *The map* $f_V$ *is sometimes called* $V$*'s* structural function*, and the set* $\mathcal{F}$ *is called a set of structural functions for* $\mathcal{V}$*.*

- $\mathcal{A}$ *is the* valuation *function, assigning to every* $X \in \mathcal{U} \cup \mathcal{V}$ *a value* $\mathcal{A}(X) \in \mathcal{R}(X)$*. For each endogenous variable, the valuation should* comply with *the variable's structural function. In other words, for every* $V_j \in \mathcal{V}$*, the following should hold:*

$$\mathcal{A}(V_j) = f_{V_j}\Big(\mathcal{A}(U_1), \ldots, \mathcal{A}(U_m), \mathcal{A}(V_1), \ldots, \mathcal{A}(V_{j-1}), \mathcal{A}(V_{j+1}), \ldots, \mathcal{A}(V_n)\Big).$$

In a causal model $\langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$, the functions in $\mathcal{F}$ describe the causal relationship between the variables. Using these functional dependencies, we can define what it means for a variable to directly causally affect another variable.[4]

[3] Given $(X_1, \ldots, X_k) \in (\mathcal{U} \cup \mathcal{V})^k$, abbreviate $\mathcal{R}X_1 \times \cdots \times \mathcal{R}(X_k)$ as $\mathcal{R}(X_1, \ldots, X_k)$.

[4] This notion of a *direct cause* is adopted from Galles and Pearl [1998a]; it is related to the notion of a variable having a *direct effect* on another, as discussed in Pearl [2000] in the context of Causal Bayes Nets. The notions defined here differ from Halpern's notion of *affect* Halpern [2000], and this affects the axiomatization: axiom HP6 (Table 2.5) has the same function as C6 in Halpern [2000] (ensuring that the canonical model is recursive), but does so in a slightly different way.

**2.** Definition (Causal dependency). *Let $\mathcal{F}$ be a set of structural functions for $\mathcal{V}$. Given an endogenous variable $V_j \in \mathcal{V}$, rename each other variable in $\mathcal{S}$, the variables $U_1, \ldots, U_m, V_1, \ldots, V_{j-1}, V_{j+1}, \ldots, V_n$, as $X_1, \ldots, X_{m+n-1}$, respectively.*

*We say that, under the structural functions in $\mathcal{F}$, an endogenous variable $V_j \in \mathcal{V}$ is* directly causally affected *by a variable $X_i \in (\mathcal{U} \cup \mathcal{V}) \setminus \{V_j\}$ (in symbols, $X_i \hookrightarrow_{\mathcal{F}} V_j$) if and only if there is a tuple*

$$(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{m+n-1}) \in \mathcal{R}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{m+n-1})$$

*and there are $x_i' \neq x_i'' \in \mathcal{R}(X_i)$ such that*

$$f_{V_j}(x_1, \ldots, x_i', \ldots, x_{m+n-1}) \neq f_{V_j}(x_1, \ldots, x_i'', \ldots, x_{m+n-1}).$$

*When $X_i \hookrightarrow_{\mathcal{F}} V_j$, we will also say that $X_i$ is a* causal parent *of $V_j$. The relation $\hookrightarrow_{\mathcal{F}}^{+}$ is the transitive closure of $\hookrightarrow_{\mathcal{F}}$.*

As it is common in the literature, we restrict ourselves to causal models in which circular causal dependencies do not occur.

**3.** Definition (Recursive causal model). *A set of structural functions $\mathcal{F}$ is* recursive *if and only if $\hookrightarrow_{\mathcal{F}}^{+}$ is a strict partial order (i.e., an asymmetric [hence irreflexive] and transitive relation, so there are no cycles). A causal model $\langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$ is* recursive *if and only if $\mathcal{F}$ is recursive. In this text, a recursive causal model will be called simply a* causal model.

The most important notion of this formalisation of causal reasoning is that of an *intervention*. This notion refers to the action of changing the values of variables in the system. Before we define an intervention formally, let us first introduce the notion of assignment.

**4.** Definition (Assignment). *Let $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ be a signature. An* assignment *on $\mathcal{S}$ is an expression $\overrightarrow{X}{=}\overrightarrow{x}$ where $\overrightarrow{X}$ is a tuple of different variables in $\mathcal{U} \cup \mathcal{V}$ (that is, $\overrightarrow{X} = (X_1, \ldots, X_k) \in (\mathcal{U} \cup \mathcal{V})^k$ for some $k \in \mathbb{N}$, with $X_i \neq X_j$ for $i \neq j$), and $\overrightarrow{x} \in \mathcal{R}(\overrightarrow{X})$.*

Now, an intervention that sets a variable $X$ to the value $x$ can be defined as an operation that maps a given model $M$ to a new model $M_{X=x}$, which is the same except that the function determining the value of $X$ is replaced by the constant function mapping $X$ to $x$. In other words, $X$ is cut off from all its causal dependencies and fixed to the value $x$.

**5.** Definition (Intervention). *Let $M = \langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$ be a causal model; let $\overrightarrow{X}{=}\overrightarrow{x}$ be an assignment on $\mathcal{S}$. The causal model $M_{\overrightarrow{X}=\overrightarrow{x}} = \langle \mathcal{S}, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}} \rangle$, resulting from an intervention setting the values of variables in $\overrightarrow{X}$ to $\overrightarrow{x}$, is such that*

- $\mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}$ *is as* $\mathcal{F}$ *except that, for each endogenous variable* $X_i$ *in* $\overrightarrow{X}$*, the function* $f_{X_i}$ *is replaced by a* constant *function* $f'_{X_i}$ *that returns the value* $x_i$ *regardless of the values of all other variables.*
- $\mathcal{A}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}$ *is the unique valuation where* ***(i)*** *the value of each exogenous variable not in* $\overrightarrow{X}$ *is exactly as in* $\mathcal{A}$*,* ***(ii)*** *the value of each each exogenous variable* $X_i$ *in* $\overrightarrow{X}$ *is the provided* $x_i$*, and* ***(iii)*** *the value of each endogenous variable complies with its* new *structural function (that in* $\mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}$*).*[5]

Building on the notion of intervention we can now extend a propositional language with a new type of sentence that considers the consequences of interventions. The expression $[\overrightarrow{X}{=}\overrightarrow{x}]\gamma$ should be read as the counterfactual conditional *if the variables in* $\overrightarrow{X}$ *were set to the values* $\overrightarrow{x}$*, respectively, then* $\gamma$ *would be the case.*

**6.** Definition. *Formulas* $\phi$ *of the language* $\mathcal{L}_C$ *based on the signature* $\mathcal{S}$ *are given by*

$$\begin{array}{ll} \gamma ::= Z{=}z \mid \neg\gamma \mid \gamma \wedge \gamma & \text{for } Z \in \mathcal{U} \cup \mathcal{V} \text{ and } z \in \mathcal{R}(Z) \\ \phi ::= Z{=}z \mid \neg\phi \mid \phi \wedge \phi \mid [\overrightarrow{X}{=}\overrightarrow{x}]\gamma & \text{for } \overrightarrow{X}{=}\overrightarrow{x} \text{ an assignment on } \mathcal{S} \end{array}$$

The language makes free use of Boolean operators, but it forbids the nesting of intervention operators $[\overrightarrow{X}{=}\overrightarrow{x}]$ (see Briggs [2012] for a way to extend the system with nested interventions). Formulas of $\mathcal{L}_C$ are evaluated in causal models $\langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$. The semantic interpretation for Boolean operators is the usual; for the rest,

$$\begin{array}{lll} \langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle \models Z{=}z & \text{iff} & \mathcal{A}(Z) = z \\ \langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle \models [\overrightarrow{X}{=}\overrightarrow{x}]\gamma & \text{iff} & \langle \mathcal{S}, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}} \rangle \models \gamma \end{array}$$

## 2.3 Limitations of the standard system

The notion of a causal model contains an incredible amount of extra information compared to classical models. Not only does it tell us which variables depend causally on which other variables, but it also determines the exact character of this dependence. On the side of the language this wealth of information is

[5] Note that, since $\mathcal{F}$ is recursive, the valuation $\mathcal{A}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}$ is uniquely determined. First, the value of every exogenous variable $U$ is uniquely determined, either from $\overrightarrow{x}$ (if $U$ occurs in $\overrightarrow{X}$) or else from $\mathcal{A}$ (if $U$ does not occur in $\overrightarrow{X}$). Second, the value of every endogenous variable $V$ is also uniquely determined, either from $\overrightarrow{x}$ (if $V$ occurs in $\overrightarrow{X}$, as $V$'s new structural function is a constant) or else from the (recall: recursive) structural functions in $\mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}$ (if $V$ does not occur in $\overrightarrow{X}$).

then explored in terms of counterfactual conditionals using the concept of an intervention. This is where the actual causal reasoning happens. The standard logic of causal reasoning is in fact a logic of counterfactual reasoning. This is no accident: Judea Pearl, founder of the approach to causal reasoning introduced above, sees both concepts as intimately related. He argues that only when an agent can evaluate counterfactual conditionals does she fully engage with causal reasoning Pearl [2009], Pearl and Mackenzie [2019]. Counterfactual reasoning *is* the highest level of causal reasoning – a level that even the most advanced AI technology doesn't even come close to.[6]

Still, the basic causal framework has some limitations. An important one is that causal (or counterfactual) reasoning does not stand on its own: it does interact with other forms of reasoning. For instance, and as we illustrated in the introduction, counterfactual reasoning also considers the effect interventions have on the epistemic state of (observing) agents. We can reason that *If Peter had pushed the button, he would have known that his flashlight is broken*, which involves thinking about Peter's epistemic state after observing a causal intervention. This type of reasoning allows us to plan our actions (try out a flashlight before we take it for a night walk), and also influences our interaction with other agents (if you want Peter to come back from his walk, you should tell him to test his flashlight before he leaves). Therefore, a full account of the logic of causal reasoning needs to model its interaction with epistemic reasoning as well. The next section takes a first step in this direction: it adds a representation of the epistemic state of an agent to the model, extending the language with expressions that can talk about knowledge and knowledge-update in the context of causal reasoning.

There is another perspective from which such an epistemic extension of the standard framework can be motivated. In recent years there has been growing interest in the logic of dependence/determinacy. For instance, the IF logic of Mann et al. [2011] expresses dependence by decorations of the quantifiers. Then, Väänänen [2007] and Baltag and van Benthem [2020] use a primitive expression indicating that the value of one variable depends on that of another. In all these cases, the discussed notion of dependence/determinacy relies on considering a multiplicity of valuations in the model: the variable $Y$ depends on (it is determined by) the variables $X_1, \ldots, X_n$ when, in all valuations that are being considered, fixing the value of the latter also fixes the value of the former. This gives rise to the question of how the notion of causal dependence modelled by the just introduced framework interacts with the notions of dependence/determinacy modelled by these alternative frameworks,

[6] The other two levels that Pearl distinguishes are the level of association, which is based on observation, and the level of intervention, which is based on doing. Modern AI technology is for him still at the first level: association. Counterfactual reasoning is not possible without a true understanding of *why* things happen – in our terminology, it is not possible without knowing the causal relationships as determined by $\mathcal{F}$.

and how causal dependence fits into a general picture of reasoning with and about dependencies. Interestingly, extending the standard causal reasoning approach with basic epistemic notions gives us another way to express the same notion of dependence as studied in the works just cited. This, then, allows us to compare different notions of dependency within one logical system. We will come back to this connection in Section 2.6.

## 2.4 Epistemic causal models

The first step towards a framework that combines causal with epistemic reasoning is adding a representation of the epistemic state of an agent to the causal model. This is done by adding a set of valuations $\mathcal{T}$, representing the alternatives the agent considers possible.

**7.** Definition (Epistemic causal model). *An* epistemic (note: recursive) causal model *is a tuple* $\langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle$ *where* $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ *is a signature,* $\mathcal{F}$ *is a (note: recursive) set of structural functions for* $\mathcal{V}$*, and* $\mathcal{T}$ *is a non-empty set of valuation functions for* $\mathcal{U} \cup \mathcal{V}$*, each one of them complying with* $\mathcal{F}$*.*

As a consequence of this definition, it is not possible to model uncertainty about the causal dependencies.

Investigating the consequences of lifting this restriction is left for future research. The next step is to extend the notion of an intervention to epistemic causal models.

**8.** Definition (Intervention). *Let* $\mathrm{E} = \langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle$ *be an epistemic causal model; let* $\overrightarrow{X}{=}\overrightarrow{x}$ *be an assignment on* $\mathcal{S}$*. The epistemic causal model* $\mathrm{E}_{\overrightarrow{X}=\overrightarrow{x}} = \langle \mathcal{S}, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}} \rangle$*, resulting from an intervention setting the values of variables in* $\overrightarrow{X}$ *to* $\overrightarrow{x}$*, is such that*

- $\mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}$ *is defined from* $\mathcal{F}$ *just as in Definition 7,*
- $\mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}} := \{\mathcal{A}'^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}} \mid \mathcal{A}' \in \mathcal{T}\}$ *(see Definition 7.*

*Note how* $\langle \mathcal{S}, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}} \rangle$ *is indeed an epistemic causal model, as* $\mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}$ *is recursive and all valuations in* $\mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}$ *comply with it.*

This definition assumes that the agent has full epistemic access to the effect of the intervention on the model; thus, she knows that the intervention takes place (in the counterfactual scenario considered). This makes a lot of sense if you think of the agent whose epistemic state is modelled as the one engaging in the counterfactual thinking. It is less plausible in connection to counterfactual thinking about the knowledge states of other agents. But this is something

that we can leave for now, as we will not consider epistemic causal models for multiple agents in this paper.

Based on these changes on the semantic side, we can now extend the object language with expressions that talk about the epistemic state of the agent. More specifically, we add the operator $K$ for knowledge and "!" for information update. In other words, we understand "!" as expressing the action of observing or receiving information.

**9.** Definition. *Formulas $\phi$ of the language $\mathcal{L}_{PAKC}$ based on $\mathcal{S}$ are given by*

$$\begin{array}{ll} \gamma ::= Z{=}z \mid \neg\gamma \mid \gamma \wedge \gamma \mid K\gamma \mid [\gamma!]\gamma & \text{for } Z \in \mathcal{U} \cup \mathcal{V} \text{ and } z \in \mathcal{R}(Z) \\ \phi ::= Z{=}z \mid \neg\phi \mid \phi \wedge \phi \mid K\phi \mid [\phi!]\phi \mid [\vec{X}{=}\vec{x}]\gamma & \text{for } \vec{X}{=}\vec{x} \text{ an assignment on } \mathcal{S} \end{array}$$

Other Boolean operators ($\vee, \rightarrow, \leftrightarrow$) can be defined as usual. Note how, although the language makes free use of Boolean, epistemic and announcement operators ($K$ and $[\phi!]$, for the latter two), nested intervention is again not allowed.[7] Note also how the tuple vector $\vec{X}$ can be empty, in which case $[\vec{X}{=}\vec{x}]\gamma$ becomes $\gamma$. The semantics for this extended language is straightforward.

**10.** Definition. *Formulas of $\mathcal{L}_{PAKC}$ are evaluated in a pairs $(\mathrm{E}, \mathcal{A})$ with $\mathrm{E} = \langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle$ an epistemic causal model and $\mathcal{A} \in \mathcal{T}$. The semantic interpretation for Boolean operators is the usual; for the rest,*

$$\begin{array}{lll} (\mathrm{E}, \mathcal{A}) \models Z{=}z & \textit{iff} & \mathcal{A}(Z) = z \\ (\mathrm{E}, \mathcal{A}) \models K\phi & \textit{iff} & (\mathrm{E}, \mathcal{A}') \models \phi \textit{ for every } \mathcal{A}' \in \mathcal{T} \\ (\mathrm{E}, \mathcal{A}) \models [\psi!]\phi & \textit{iff} & (\mathrm{E}, \mathcal{A}) \models \psi \textit{ implies } (\mathrm{E}^{\psi}, \mathcal{A}) \models \phi \\ (\mathrm{E}, \mathcal{A}) \models [\vec{X}{=}\vec{x}]\gamma & \textit{iff} & (\mathrm{E}_{\vec{X}=\vec{x}}, \mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}) \models \gamma \end{array}$$

*with $\mathrm{E}^{\psi} = \langle \mathcal{S}, \mathcal{F}, \mathcal{T}^{\psi} \rangle$ such that $\mathcal{T}^{\psi} := \{\mathcal{A}' \in \mathcal{T} \mid (\mathrm{E}, \mathcal{A}') \models \psi\}$. Note how $\mathrm{E}^{\psi}$ is an epistemic causal model: $\mathcal{F}$ is recursive, and all valuations in $\mathcal{T}^{\psi}$ comply with it.*

Finally, we define an operator $\rightsquigarrow$ in terms of the existing vocabulary as a way to express causal dependency in the object language.

**11.** Definition. *Take $X$ and $Z$ in $\mathcal{U} \cup \mathcal{V}$. The formula $X \rightsquigarrow Z$ is defined as*

$$\bigvee_{\substack{\vec{w} \in \mathcal{R}((\mathcal{U} \cup \mathcal{V}) \setminus \{X, Z\}), \\ \{x_1, x_2\} \subseteq \mathcal{R}(X), x_1 \neq x_2, \\ \{z_1, z_2\} \subseteq \mathcal{R}(Z), z_1 \neq z_2}} [\vec{W}{=}\vec{w}, X{=}x_1]Z{=}z_1 \ \wedge \ [\vec{W}{=}\vec{w}, X{=}x_2]Z{=}z_2,$$

[7] However, notice that the semantics already allows for nested occurrences of all dynamic operators. We will extend the proofs of sound- and completeness to the unrestricted language in the future.

A formula $X \rightsquigarrow Z$ should be read as *"$X$ has a direct causal effect on $Z$"*. It holds when there is a vector $\vec{w}$ of values for variables in $\mathcal{R}(\mathcal{U} \cup \mathcal{V} \setminus \{X, V\})$ and two different values $x_1, x_2$ for $X$ that produce two different values $z_1, z_2$ for $Z$ (cf. Halpern [2000]). When $Z \in \mathcal{V}$, it is clear that $\rightsquigarrow$ is the syntactic counterpart of the relation "$\hookrightarrow$" of Definition 2.

## 2.5 Axiomatization

*Propositional*:

P: $\phi$ for $\phi$ an instance of a tautology

MP: From $\phi \rightarrow \psi$ and $\phi$ derive $\psi$

*Intervention*:

HP1: $[\vec{X}{=}\vec{x}]Z{=}z \rightarrow \neg[\vec{X}{=}\vec{x}]Z{=}z'$ for $z \neq z' \in \mathcal{R}(Z)$

HP2: $\bigvee_{z \in \mathcal{R}(Z)} [\vec{X}{=}\vec{x}]Z{=}z$

HP3: $\left([\vec{X}{=}\vec{x}]Z{=}z \wedge [\vec{X}{=}\vec{x}]W{=}w\right) \rightarrow [\vec{X}{=}\vec{x}, Z{=}z]W{=}w$

HP4: $[\vec{X}{=}\vec{x}, Z{=}z]Z{=}z$

HP5: $\left([\vec{X}{=}\vec{x}, Z{=}z]W{=}w \wedge [\vec{X}{=}\vec{x}, W{=}w]Z{=}z\right) \rightarrow [\vec{X}{=}\vec{x}]W{=}w$ for $W \neq Z$

HP6: $(Z_0 \rightsquigarrow Z_1 \wedge \cdots \wedge Z_{k-1} \rightsquigarrow Z_k) \rightarrow \neg(Z_k \rightsquigarrow Z_0)$

RH1: $[\vec{X}{=}\vec{x}](\phi \wedge \psi) \leftrightarrow ([\vec{X}{=}\vec{x}]\phi \wedge [\vec{X}{=}\vec{x}]\psi)$

RH2: $[\vec{X}{=}\vec{x}]\neg\phi \leftrightarrow \neg[\vec{X}{=}\vec{x}]\phi$

EX: $U{=}u \leftrightarrow [\vec{X}{=}\vec{x}]U{=}u$ for $U \in \mathcal{U}$ with $U \notin \vec{X}$

*Epistemic*:

K: $K(\phi \rightarrow \psi) \rightarrow (K\phi \rightarrow K\psi)$

N: From $\phi$ derive $K\phi$

T: $K\phi \rightarrow \phi$

4: $K\phi \rightarrow KK\phi$

5: $\neg K\phi \rightarrow K\neg K\phi$

*Epistemic+Intervention*:

CM: $[\vec{X}{=}\vec{x}]K\phi \leftrightarrow K[\vec{X}{=}\vec{x}]\phi$

*Announcement*:

RP1: $[\psi!][\vec{X}{=}\vec{x}]Z{=}z \leftrightarrow (\psi \rightarrow [\vec{X}{=}\vec{x}]Z{=}z)$

RP2: $[\psi!]\neg\phi \leftrightarrow (\psi \rightarrow \neg[\psi!]\phi)$

RP3: $[\psi!](\phi \wedge \chi) \leftrightarrow ([\psi!]\phi \wedge [\psi!]\chi)$

RP4: $[\psi!]K\phi \leftrightarrow (\psi \rightarrow K(\psi \rightarrow [\psi!]\phi))$

RE: From $\psi_1 \leftrightarrow \psi_2$ derive $\phi \leftrightarrow \phi[\psi_2/\psi_1]$, with $\phi[\psi_2/\psi_1]$ a formula obtained by replacing one or more non-announcement occurrences of $\psi_1$ in $\phi$ with $\psi_2$.[8]

Table 2.1: Axiom system $\mathsf{L}_{PAKC}$

The axiom system $\mathsf{L}_{PAKC}$ is presented in Table 2.5.The *intervention* axioms, HP1-HP6, RH1 and RH2, are the standard axiomatization for the intervention operator over recursive causal models, with EX an additional axiom indicating that an exogenous variable is immune to interventions to any other variables. Then, the *epistemic* part contains the standard modal S5 axiomatization for truthful knowledge with positive and negative introspection.

Axiom CM indicates that what the agent will know after an intervention ($[\overrightarrow{X}{=}\overrightarrow{x}]K\phi$) is exactly what she knows now about the effects of the intervention ($K[\overrightarrow{X}{=}\overrightarrow{x}]\phi$). Although maybe novel in the literature on causal models, the axiom is simply an instance of the more general DEL pattern of interaction between knowledge and a deterministic action without precondition. Finally, axioms RP2-RP4 and rule RE in the *announcement* part are a *reduction-based* axiomatisation for public announcements in the DEL style. Here, axioms RP4 and RP1 are the most important. The first, RP4, is the well-known reduction axiom for announcement and knowledge, stating that knowing $\phi$ after an announcement of $\psi$ is equivalent to knowing, conditionally on $\psi$, that the announcement of $\psi$ would make $\phi$ true.[9] The second, RP1, establishes the reduction for 'atoms' of the form $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$; when $\overrightarrow{X}$ is not empty, it states that a public announcement does not change the causal rules in the model.

The axiom system $\mathsf{L}_{PAKC}$ is sound and complete for $\mathcal{L}_{PAKC}$ in epistemic causal models. Here is the argument for soundness.

**1.** THEOREM. *The axiom system $\mathsf{L}_{PAKC}$ is sound for $\mathcal{L}_{PAKC}$ in epistemic causal models.*

*Proof.* For the soundness of HP1-HP6, RH1 and RH2 on causal models (enough for soundness on epistemic causal models, as evaluating the formulas does not require a change in valuation), see Halpern [2000]. For the soundness of K, N, T, 4, and 5 on relational structures with an equivalence relation (equivalent to having a simple set of epistemic alternatives, as epistemic causal models have), see Fagin et al. [1995], Blackburn et al. [2001]. For the soundness of RP1-RP4 when $[\psi!]$ describes the effect of a deterministic domain-reducing model operation, see Wang and Cao [2013].

For axioms EX and CM, take any $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}\rangle, \mathcal{A})$. For EX note how, for any $\overrightarrow{X}{=}\overrightarrow{x}$, the valuations $\mathcal{A}$ and $\mathcal{A}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}$ assign the same value to *exogenous* variables not occurring in $\overrightarrow{X}$ (Definition 5). For CM, note how ***(i)*** $K[\overrightarrow{X}{=}\overrightarrow{x}]\phi$ holds at $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}\rangle, \mathcal{A})$ iff $\phi$ holds at $(\langle \mathcal{S}, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}\rangle, \mathcal{A}'^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}})$ for every $\mathcal{A}' \in \mathcal{T}$, and ***(ii)*** $[\overrightarrow{X}{=}\overrightarrow{x}]K\phi$ holds at $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}\rangle, \mathcal{A})$ iff $\phi$ holds at $(\langle \mathcal{S}, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}\rangle, (\mathcal{A}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}})')$

[8] A non-announcement occurrence of $\psi$ in $\phi$ is an occurrence of $\psi$ in $\phi$ where $\psi$ is not inside the *brackets* of an announcement operator.

[9] Note how the announcement of $\psi$ is a deterministic action *with precondition* $\psi$. Hence the similarities and differences between RP4 and CM.

for every $(\mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}})' \in \mathcal{T}^{\mathcal{F}}_{\vec{X}=\vec{x}}$. Then it is enough to notice how, by Definition 8, the set of relevant valuations for the second, $\mathcal{T}^{\mathcal{F}}_{\vec{X}=\vec{x}}$, is exactly the set of relevant valuations for the first, $\{\mathcal{A}'^{\mathcal{F}}_{\vec{X}=\vec{x}} \mid \mathcal{A}' \in \mathcal{T}\}$. Finally, soundness of RE follows from two facts: the truth-value of every formula depends on the truth-value of its subformulas, and model operations (intervention and announcements) produce epistemic causal models. Thus, substituting a non-announcement subformula for a formula that is semantically equivalent in the given class of structures does not affect the final result.

The argument for completeness uses two steps. ***(i)*** First, using the reduction axioms technique, it will be shown that $\mathsf{L}_{PAKC}$ allow us to translate any formula in $\mathcal{L}_{PAKC}$ into a logically equivalent one without public announcements. ***(ii)*** Then, relying on the canonical model construction for both causal models Halpern [2000] and epistemic models [Fagin et al., 1995, Chapter 3], it will be shown that $\mathsf{L}_{PAKC}$ is complete for the language without public announcements. For the full proof, see the appendix of Chapter 2.

**2.** Theorem. *The axiom system $\mathsf{L}_{PAKC}$ is complete for $\mathcal{L}_{PAKC}$ in epistemic causal models.*

## 2.6 Discussion

In this section we will compare our proposal to the Causal Team Semantics developed in Barbero and Sandu [2019, 2020], Barbero and Yang. Causal Team Semantics was proposed with the intention of supporting languages that discuss both accidental and causal dependencies. This is a topic that has gained quite some interest in recent years; see, for instance, also (Chockler and Halpern [2004], Ibeling and Icard [2020]). Causal Team Semantics was developed along the lines of a non-modal tradition of logics of dependence and independence (e.g. Väänänen [2007], Mann et al. [2011]) by extending the so-called *team semantics* (Hodges [1997]) with elements taken from causal inference. Even though the focus there is not on combining causal with epistemic reasoning, this framework bears many similarities to the one we are using, which is why we will discuss it here in detail. Furthermore, this also allows us to say a bit more on the topic of dependence from the perspective of our proposal.

Let us quickly introduce the central notions of Causal Team Semantics to facilitate a comparison of the two frameworks. A causal team[10] is a tuple $T = \langle \mathcal{T}, \mathcal{F} \rangle$ where $\mathcal{F}$ is defined similarly as in our paper[11] and $\mathcal{T}$ is a possibly

[10] We are presenting here the definition from Barbero and Yang, which, save for implementation details, corresponds to what are called *fully defined* causal teams in Barbero and Sandu [2019] (where a more general notion is considered).

[11] With some additional machinery (which is not worth exploring here) to keep track of the domains of the functions. For simplicity, we may assume here that $\mathcal{F}$ is defined in the same way as for causal epistemic models.

empty set of valuations that comply with $\mathcal{F}$. Papers on Causal Team Semantics consider a variety of languages. The focus here is the one we shall call $\mathcal{L}_{COD}$, which is similar to the standard causal language (thus allowing to express various notions of causal dependence in terms of counterfactuals) except for the additional *dependence atoms* "$=(X_1, ..., X_n; Y)$", which expresses (accidental) dependency of the variable $Y$ on the variables $X_1$ to $X_n$. A sentence $=(X_1, ..., X_n; Y)$ is interpreted as the claim that any two states $s$ and $s'$ that agree on the valuation of the variables $X_1, ..., X_n$ also have to agree on the value they assign to $Y$. Below the complete semantics of $\mathcal{L}_{COD}$ is given, using the notation of this manuscript.[12]

| | | |
|---|---|---|
| $T \models Y{=}y$ | iff | $s(Y) = y$ for all $s \in \mathcal{T}$ |
| $T \models Y{\neq}y$ | iff | $s(Y) \neq y$ for all $s \in \mathcal{T}$ |
| $T \models =(X_1, ..., X_n; Y)$ | iff | for all $s, s' \in \mathcal{T}$, if $s(X_i) = s'(X_i)$ for $1 \leqslant i \leqslant n$, then $s(Y) = s'(Y)$ |
| $T \models \phi \wedge \psi$ | iff | $T \models \phi$ and $T \models \psi$ |
| $T \models \phi \vee \psi$ | iff | there are $\mathcal{T}_1 \cup \mathcal{T}_2 = \mathcal{T}$ such that $\langle \mathcal{T}_1, \mathcal{F} \rangle \models \phi$ and $\langle \mathcal{T}_2, \mathcal{F} \rangle \models \psi$ |
| $T \models \alpha \supset \psi$ | iff | $\langle \mathcal{T}^{\alpha}, \mathcal{F} \rangle \models \psi$, for $\mathcal{T}^{\alpha} := \{s \in \mathcal{T} \mid (\{s\}, \mathcal{F}) \models \alpha\}$ and $\alpha$ without dependence atoms |
| $T \models \overrightarrow{X}{=}\overrightarrow{x} \boxdot\!\!\rightarrow \psi$ | iff | $\langle \mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}} \rangle \models \psi$, with $\mathcal{T}^{\mathcal{F}}_{\overrightarrow{X}=\overrightarrow{x}}$ and $\mathcal{F}_{\overrightarrow{X}=\overrightarrow{x}}$ as in Definition 8. |

Notice that formulas are evaluated on a causal team globally, i.e., no valuation in $\mathcal{T}$ is isolated as being 'the actual world'. Due to this global perspective, the interpretation of some connectives ($\vee$ and $\supset$) differs from that given on epistemic causal models; however, these connectives behave classically if applied to subformulas without occurrences of dependence atoms, and also when $\mathcal{T}$ is a singleton.

From their definitions, it is clear that an epistemic causal and a causal team are identical objects; the only difference is that, for evaluating formulas, the former requires an 'actual world'. On the syntactic side, even though the truth clauses of the logical operators differ in various respects, we can find several equivalences. For instance, the notion of dependence from team semantics can be expressed in our formal language as well.[13] Indeed, interpret the object $\mathcal{T}$ of a causal team as the epistemic state of some agent. Then, the statement $Y = y$ of causal team semantics can be understood as a claim about the knowledge of the agent, written in our language as $K(Y = y)$. Building on this translation, we can express that variable $Y$ depends on the variables $\overrightarrow{X}$ as the following claim:

[12] A full definition of the syntax can be found in the appendix of Chapter 2

[13] As far as we know, this has been first observed, independently, in van Eijck et al. [2017] and Baltag [2016], in the context of epistemic languages with modalities for the knowledge of values.

for all possible valuations $\overrightarrow{x}$ of $\overrightarrow{X}$ there is some value $y$ of $Y$ such that the agent knows that if she would observe $\overrightarrow{X} = \overrightarrow{x}$, she would know that $Y$ has value $y$.

$$\bigwedge_{\overrightarrow{x} \in \mathcal{R}(\overrightarrow{X})} \bigvee_{y \in \mathcal{R}(Y)} [(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y{=}y).$$

With this idea in mind we can define a translation of the non-nested formulas of $\mathcal{L}_{COD}$.[14] Setting aside for a moment the case of the operator $\supset$, and using $A$ to denote the set of all possible valuations for $\mathcal{U} \cup \mathcal{V}$, the translation is given by the following clauses.

$$tr(Y{=}y) := K(Y{=}y) \qquad tr(\phi_1 \wedge \phi_2) := tr(\phi_1) \wedge tr_(\phi_2)$$

$$tr(Y{\neq}y) := K(\neg(Y{=}y)) \qquad tr(\overrightarrow{X}{=}\overrightarrow{x} \boxdot\!\!\rightarrow \phi) := [\overrightarrow{X}{=}\overrightarrow{x}]tr(\phi)$$

$$tr(\phi \vee \psi) := \bigvee_{S \subseteq A} K\Big([(\bigvee_{\overrightarrow{Y}=\overrightarrow{y} \in S} \overrightarrow{Y} = \overrightarrow{y})!]tr(\phi) \ \wedge \ [(\neg \bigvee_{\overrightarrow{Y}=\overrightarrow{y} \in S} \overrightarrow{Y} = \overrightarrow{y})!]tr(\psi)\Big)$$

$$tr(=(X_1, ..., X_n; Y)) := \bigwedge_{\overrightarrow{x} \in \mathcal{R}(\overrightarrow{X})} \bigvee_{y \in \mathcal{R}(Y)} [(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y{=}y)$$

Formulas of the form $\alpha \supset \psi$ translate into public announcement formulas. However, in order to play the role of announcement, $\alpha$ cannot be translated using $tr$, as announcements are evaluated according to the classical meaning. We need instead a simpler translation $e$ which just replaces logical operators with their counterparts in $\mathcal{L}_{PAKC}$ ($X{\neq}x$ is replaced by $\neg(X{=}x)$; $\beta \supset \gamma$ by $\beta \rightarrow \gamma$; $\overrightarrow{X} = \overrightarrow{x} \boxdot\!\!\rightarrow \phi$ by $[\overrightarrow{X} = \overrightarrow{x}]\phi$; $\wedge$ and $\vee$ are left unaltered, or, more precisely, $\beta \vee \gamma$ is replaced by $\neg(\neg\beta \wedge \neg\gamma)$). Then we can define $tr$ for $\supset$ as follows:

$$tr(\alpha \supset \phi) := [e(\alpha)!]tr(\phi)$$

Let $T = (\mathcal{T}, \mathcal{F})$ be a causal team; let $E = \langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle$ be a causal epistemic model. The provided translation satisfies the following (for a proof see the appendix of Chapter 2).

**1.** Proposition (Global translation). *For any causal team $\langle \mathcal{T}, \mathcal{F} \rangle$ over a* finite *signature $\mathcal{S}$ and any formula $\phi \in \mathcal{L}_{COD}$, we have $\langle \mathcal{T}, \mathcal{F} \rangle \models \phi$ if and only if, for all $\mathcal{A} \in \mathcal{T}$, we have $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle, \mathcal{A}) \models tr(\phi)$.*

[14] A formula is non-nested if, in every subformula of the form $\overrightarrow{X}{=}\overrightarrow{x} \boxdot\!\!\rightarrow \phi$, no $\boxdot\!\!\rightarrow$ occurs inside $\phi$. Providing a translation for these formulas is sufficient, since every formula of the causal team language is provably equivalent to a non-nested one.

This result compares truth on a causal team with validity over an epistemic causal model. On the other hand, a different translation of the dependence atom from van Eijck et al. [2017], Baltag [2016] suggests an alternative, "local" translation. Let $tr^*$ be as $tr$, except for the following clauses (notice the additional $K$ operator in both clauses):

$$tr^*(=(X_1,...,X_n;Y)) := \bigwedge_{\vec{x}\in\mathcal{R}(\vec{X})} \bigvee_{y\in\mathcal{R}(Y)} K[(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y{=}y)$$
$$tr^*(\alpha \supset \phi) := K[e(\alpha)!]tr(\phi)$$

Now we have the following result.

**2.** Proposition (Local translation). *For any causal team $\langle \mathcal{T}, \mathcal{F}\rangle$ over a* finite *signature $\mathcal{S}$ and any formula $\phi \in \mathcal{L}_{COD}$, we have:*

***(i)*** *If $\langle \mathcal{T}, \mathcal{F}\rangle \models \phi$, then, for all $\mathcal{A} \in \mathcal{T}$, $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}\rangle, \mathcal{A}) \models tr(\phi)$.*

***(ii)*** *If* there is *an $\mathcal{A} \in \mathcal{T}$ such that $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}\rangle, \mathcal{A}) \models tr(\phi)$, then $\langle \mathcal{T}, \mathcal{F}\rangle \models \phi$.*

This result shows that, *in the finite case*, $\mathcal{L}_{PAKC}$ is at least as expressive as $\mathcal{L}_{COD}$. Despite this, the way the notion of (accidental) dependence is spelled out in the two languages differs in an interesting way. While it is a primitive element in the language of Causal Team Semantics, the way it is definable in our epistemic framework emphasises what we can *do* with such a concept of dependence: we can make predictions based on what we observe. Furthermore, it is interesting to notice the similarity between this translation of (accidental) dependence and the way causal dependence is expressed. It is also not defined as a primitive in the language, but can be expressed using counterfactuals, which work based on the concept of intervention. These counterfactuals, in turn, focus on what you can do with causal information: prediction based on intervention.

Based on the counterfactual expression, various notions of causal dependence can be defined. We saw one already in Section 4.5, Definition 11: $X \rightsquigarrow Z$, which expresses that $X$ is a causal parent of $Z$ (if $Z$ is an endogenous variable). The local translation of the notion of dependence from Causal Team Semantics into our framework suggests a different notion of causal dependence. We repeat the local translation below under the name of e-dependence. C-dependence defines the corresponding causal notion.[15]

- $Y$ *e-depends* on $X$ in $(E, \mathcal{A})$ iff $(E, \mathcal{A}) \models \bigwedge_{x\in\mathcal{R}(X)} \bigvee_{y\in\mathcal{R}(Y)} K([(X = x)!]K(Y = y))$
- $Y$ *c-depends* on $Y$ in $(E, \mathcal{A})$ iff $(E, \mathcal{A}) \models \bigwedge_{x\in\mathcal{R}(X)} \bigvee_{y\in\mathcal{R}(Y)} [X = x]K(Y = y)$

[15] The additional $K$ operator in the definition of e-dependence is needed to deal with the fact that information update always checks first whether the information that the information state is updated with is true. This problem disappears in the case of interventions, because the formula you intervene with is *made* true in the hypothetical scenario you consider.

Given an epistemic causal model, C-dependence holds between a list of variables $X_1, \ldots, X_n$ and a variable $Y$ if any intervention fixing the value of the variables $X_1, \ldots, X_n$ also determines the value of $Y$ *within the epistemic state of the agent*. While this notion is certainly more robust than the notion of e-dependence, it still takes into account the epistemic state of the agent. The less the agent knows about the values of the variables, the more variables she needs to control to make sure that a variable $Y$ is in a particular state. If the agent knows more about the actual causal history of $Y$, she can predict the state of $Y$ already from smaller interventions. These kind of hybrid notions between causal and epistemic dependence that our framework allows to define deserve certainly some attention in future research.

## 2.7 Conclusions

In this paper we have moved some steps towards the integration of causal and epistemic reasoning, providing an adequate semantics, a language combining interventionist counterfactuals with (dynamic) epistemic operators and a sound and complete system of inference. Our deductive system models the thought of an agent reasoning about the consequences of hypothetical interventions and observations. It describes what the agent may deduce from her/his *a priori* pool of knowledge about a system of variables. It is therefore a logic of thought experiments. Going back to Example 1 that we discussed in the introduction, it allows us to account for the inference that Billie is not sure that if the button had been pushed, the sprinkler would have been working. However, the logic is not yet able to also model the second inference discussed in connection with this example: if Billie had pushed the button and saw that the sprinkler works, then she would have known that the circuit is closed. In order to account for this kind of reasoning we need to model how an agent may reason about (from her perspective) actual experiments. Things change significantly in such a setting: because of unobserved factors, the agent may fail to predict the outcome of an experiment; yet the outcome may sometimes be recovered from direct observation of the consequences of the experiment. The development of a such a framework will involve a more careful distinction between *observable* and *unobservable* variables. The resulting logic must necessarily abandon the right-to-left implication of axiom CM ($[\overrightarrow{X}=\overrightarrow{x}]K\phi \rightarrow K[\overrightarrow{X}=\overrightarrow{x}]\phi$), which expresses the fact that interventions cannot increase the knowledge of the agent.

Our framework has many points in common with the earlier causal team semantics, and we provided a translation between the two approaches. For the purpose of modeling causal reasoning, our semantics has the advantage, over causal team semantics, of encoding explicitly a notion of actual state of the world (and in particular, of actual value of variables). Actual values seem

to be crucial for the attempt of defining notions of *token causation* (Hitchcock [2001], Woodward [2003], Halpern [2016]), i.e. causation between events. In order to fully appreciate this advantage, though, we will need to consider richer languages with hybrid features that allow to explicitly refer to the actual values of variables.

Finally, in future work we plan to extend the setting to a multi-agent system. This involves considering not only different agents with potentially different knowledge, but also epistemic attitudes for groups (e.g., distributed and common knowledge) and the effect of inter-agent communication. This will bring the potential to contribute to the discussion about causal agency and the role of causation in the study of responsibility within AI (see, for instance, Baltag et al. [forthcoming]).

## 2.8 Appendix

### 2.8.1 Proof of Theorem 2

As mentioned, the argument for completeness proceeds in two steps: translating any formula in $\mathcal{L}_{PAKC}$ into a logically equivalent one without public announcements, and using the canonical model construction for both causal models Halpern [2000] and epistemic models Fagin et al. [1995] to show that $\mathsf{L}_{PAKC}$ is complete for the language without public announcements.

#### From $\mathcal{L}_{PAKC}$ to $\mathcal{L}_{KC}$

The translation of a formula in $\mathcal{L}_{PAKC}$ into a logically equivalent one without public announcement operators proceeds in two stages. First, the formula in $\mathcal{L}_{PAKC}$ is translated into a logically equivalent one where the only formulas inside the scope of intervention operators are of the form $Z{=}z$. This involves the use of axiom CM for putting epistemic operators $K$ outside the scope of interventions, and the use of axioms RP1-RP4 for eliminating public announcement operators *inside the scope of interventions*. The resulting formula is now built by the free use of Boolean operators, $K$ and $[\psi!]$ over 'atoms' of the form $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$. Then, axioms RP1-RP4 can be applied once more to eliminate *every remaining* public announcement operator.

To formalise the process, the following definitions will be useful.

**12.** Definition (Languages $\mathcal{L}_1$ and $\mathcal{L}_{KC}$).

- *Formulas $\xi$ of the language $\mathcal{L}_1$ are given by*

$$\xi ::= [\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \mid \neg\xi \mid \xi \wedge \xi \mid K\xi \mid [\xi!]\xi$$

*Thus, formulas in $\mathcal{L}_1$ (a fragment of $\mathcal{L}_{PAKC}$) are built by the free use of Boolean operators, $K$ and $[\psi!]$ over 'atoms' of the form $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$.*[16]

- *Formulas $\chi$ of the language $\mathcal{L}_{KC}$ are given by*

$$\chi ::= [\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \mid \neg\chi \mid \chi \wedge \chi \mid K\chi$$

*Thus, formulas in $\mathcal{L}_{KC}$ (a fragment of $\mathcal{L}_{PAKC}$) are then built by the free use of Boolean operators and $K$ over 'atoms' of the form $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$.*

The process consists of two stages: translating from $\mathcal{L}_{PAKC}$ into $\mathcal{L}_1$, and then from $\mathcal{L}_1$ into $\mathcal{L}_{KC}$.

**3.** Proposition. ***(i)*** *Every formula $\phi \in \mathcal{L}_{PAKC}$ is logically equivalent to a formula $\xi_\phi \in \mathcal{L}_1$. Moreover, $\phi \leftrightarrow \xi_\phi$ is derivable in $\mathsf{L}_{PAKC}$.* ***(ii)*** *Every formula $\xi \in \mathcal{L}_1$ is logically equivalent to a formula $\chi_\xi \in \mathcal{L}_{KC}$. Moreover, $\xi \leftrightarrow \chi_\xi$ is derivable in $\mathsf{L}_{PAKC}$.*

*Proof.* For ***(i)***, consider the translation $\mathrm{tr}_1 : \mathcal{L}_{PAKC} \to \mathcal{L}_1$ given by

$$\begin{aligned}
\mathrm{tr}_1(Z{=}z) &:= Z{=}z \\
\mathrm{tr}_1(\neg\phi) &:= \neg\,\mathrm{tr}_1(\phi) \\
\mathrm{tr}_1(\phi_1 \wedge \phi_2) &:= \mathrm{tr}_1(\phi_1) \wedge \mathrm{tr}_1(\phi_2) \\
\mathrm{tr}_1(K\phi) &:= K\,\mathrm{tr}_1(\phi) \\
\mathrm{tr}_1([\phi'!]\phi) &:= [\phi'!]\,\mathrm{tr}_1(\phi)
\end{aligned}$$

$$\begin{aligned}
\mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z) &:= [\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \\
\mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}]\neg\gamma) &:= \mathrm{tr}_1(\neg[\overrightarrow{X}{=}\overrightarrow{x}]\gamma) \\
\mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}](\gamma_1 \wedge \gamma_2)) &:= \mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}]\gamma_1 \wedge [\overrightarrow{X}{=}\overrightarrow{x}]\gamma_2) \\
\mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}]K\gamma) &:= \mathrm{tr}_1(K[\overrightarrow{X}{=}\overrightarrow{x}]\gamma) \\
\mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}][\gamma'!]Z{=}z) &:= \mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}](\gamma' \to Z{=}z)) \\
\mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}][\gamma'!]\neg\gamma) &:= \mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}](\gamma' \to \neg[\gamma'!]\gamma)) \\
\mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}][\gamma'!](\gamma_1 \wedge \gamma_2)) &:= \mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}]([\gamma'!]\gamma_1 \wedge [\gamma'!]\gamma_2)) \\
\mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}][\gamma'!]K\gamma) &:= \mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}](\gamma' \to K(\gamma' \to [\gamma'!]\gamma))) \\
\mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}][\gamma'!][\gamma'']\gamma) &:= \mathrm{tr}_1([\overrightarrow{X}{=}\overrightarrow{x}][\gamma'!]\,\mathrm{tr}_1([\gamma'']\gamma))
\end{aligned}$$

From the cases defined in the second column, it should be clear that $\mathrm{tr}_1$ does yield formulas in $\mathcal{L}_1$. Indeed, the second and third cases push intervention operators $[\overrightarrow{X}{=}\overrightarrow{x}]$ through Boolean operators until the formula directly in front of $[\overrightarrow{X}{=}\overrightarrow{x}]$ is either $Z{=}z$, or else $K$ or else $[\gamma']$. Then, while the fourth case in the second column takes $K$ outside the scope of $[\overrightarrow{X}{=}\overrightarrow{x}]$, cases six through eight 'push' $[\gamma']$ inside the formula until it has only an atom $Z{=}z$ in front, at

[16] Recall that $Z{=}z$ is the particular case of $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$ where $\overrightarrow{X}$ is empty.

which moment $[\gamma']$ is eliminated (fifth case).[17] The ninth case deals with nested announcements following an 'inside-first' strategy.

Then, note how $\models \phi \leftrightarrow \mathrm{tr}_1(\phi)$ holds for every $\phi \in \mathcal{L}_{PAKC}$. This can be shown by induction on $\phi$, with the crucial cases being those corresponding to the definitions in the second column. The first is obvious. The second and third follow from the validity of axioms RH1 and RH2, and the fourth follows from CM. Cases fifth through eighth rely on the validity of axioms RP1 through RP4, and the ninth case uses the rule RE. This last rule is used through all the cases, allowing us to replace sub-formulas for logically equivalent ones.

Finally note how, within the axiom system $\mathsf{L}_{PAKC}$, there is a derivation of $\phi \leftrightarrow \mathrm{tr}_1(\phi)$, as every non-trivial equivalence that is used for defining the translation (axioms RH1, RH2, CM, RP1-RP4 and rule RE) is in $\mathsf{L}_{PAKC}$.

For ***(ii)***, consider the translation $\mathrm{tr}_2 : \mathcal{L}_1 \to \mathcal{L}_{KC}$ given by

$$\begin{aligned}
\mathrm{tr}_2([\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z) &:= [\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z & \mathrm{tr}_2([\xi'!][\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z) &:= \mathrm{tr}_2(\xi' \to [\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z)\\
\mathrm{tr}_2(\neg\xi) &:= \neg\,\mathrm{tr}_2(\xi) & \mathrm{tr}_2([\xi'!]\neg\xi) &:= \mathrm{tr}_2(\xi' \to \neg[\xi'!]\xi)\\
\mathrm{tr}_2(\xi_1 \wedge \xi_2) &:= \mathrm{tr}_2(\xi_1) \wedge \mathrm{tr}_2(\xi_2) & \mathrm{tr}_2([\xi'!](\xi_1 \wedge \xi_2)) &:= \mathrm{tr}_2([\xi'!]\xi_1 \wedge [\xi'!]\xi_2)\\
\mathrm{tr}_2(K\xi) &:= K\,\mathrm{tr}_2(\xi) & \mathrm{tr}_2([\xi'!]K\xi) &:= \mathrm{tr}_2(\xi' \to K(\xi' \to [\xi'!]\xi))\\
& & \mathrm{tr}_2([\xi'!][\xi''!]\xi) &:= \mathrm{tr}_2([\xi'!]\,\mathrm{tr}_2([\xi''!]\xi))
\end{aligned}$$

As Wang and Cao [2013] shows, $\mathrm{tr}_2$ eliminates public announcement operators, thus yielding indeed a formula in $\mathcal{L}_{KC}$. Then, note how $\models \xi \leftrightarrow \mathrm{tr}_2(\xi)$ holds for every $\xi \in \mathcal{L}_2$. This can be shown by induction on $\chi$: the crucial cases, those corresponding to the definitions in the second column, follow from the validity of axioms RP1 through RP4. For the last entry in the second column, it is the rule RE which allow us to nest the translation function. This last rule is used through all the cases, allowing us to replace sub-formulas for logically equivalent ones. Finally note how, within the axiom system $\mathsf{L}_{PAKC}$, there is a derivation $\xi \leftrightarrow \mathrm{tr}_1(\xi)$, as every non-trivial equivalence defining the translation (axioms RP1-RP4 and rule RE) is in $\mathsf{L}_{PAKC}$.

Then,

**3.** Theorem. *Every formula $\phi \in \mathcal{L}_{PAKC}$ is logically equivalent to a formula $\chi_\phi \in \mathcal{L}_{KC}$. Moreover, $\phi \leftrightarrow \chi_\phi$ is derivable in $\mathsf{L}_{PAKC}$.*

[17] Proving that the translation ends and that announcement operators are indeed eventually eliminated requires some care. The crucial thing to notice is that, in cases sixth through eighth, the formula occurring under the scope of announcement operators on the right-hand side is less complex that the one occurring under the scope of the same announcement operator on the left-hand side. See [van Ditmarsch et al., 2008, Section 7.4] and Wang and Cao [2013] for a detailed explanation of the way the reduction works.

**Canonical model for $\mathcal{L}_{KC}$**

Now it will be shown that $\mathsf{L}_{KC}$, the fragment of $\mathcal{L}_{PAKC}$ without axioms RH1, RH2, CM and RP1-RP4, is strongly complete for $\mathcal{L}_{KC}$ over epistemic causal models. This will be done by showing, via the construction of a canonical model, that any $\mathsf{L}_{KC}$-consistent set of $\mathcal{L}_{KC}$-formulas is satisfiable in a pointed epistemic causal model. The construction here will follow those in Halpern [2000] and Fagin et al. [1995], for causal and epistemic models, respectively.

Let $\mathbb{C}$ be the set of all maximally $\mathsf{L}_{KC}$-consistent sets of $\mathcal{L}_{KC}$-formulas. The first step will be show how each $\Gamma \in \mathbb{C}$ gives raise to a causal model.

**13.** Definition (Building a causal model). *Let $\Gamma \in \mathbb{C}$ be a maximally $\mathsf{L}_{KC}$-consistent set of $\mathcal{L}_{KC}$-formulas.*

- *Let $\overrightarrow{\mathcal{U}}$ be the tuple of all exogenous variables. For each endogenous variable $V \in \mathcal{V}$, let $\overrightarrow{\mathcal{Y}}$ be the tuple of all endogenous variables in $\mathcal{V} \setminus \{V\}$. The structural function $f_V^\Gamma$ is defined, for each $\overrightarrow{u} \in \mathcal{R}(\overrightarrow{\mathcal{U}})$ and $\overrightarrow{y} \in \mathcal{R}(\overrightarrow{\mathcal{Y}})$, as*

$$f_V^\Gamma(\overrightarrow{u}, \overrightarrow{y}) = v \quad \text{if and only if} \quad [\overrightarrow{\mathcal{U}}{=}\overrightarrow{u}, \overrightarrow{\mathcal{Y}}{=}\overrightarrow{y}]V{=}v \in \Gamma$$

  *Note: axioms HP1 and HP2 ensure that $f_V^\Gamma$ is well-defined, as they guarantee $\Gamma$ has one and only one formula of the form $[\overrightarrow{\mathcal{U}}{=}\overrightarrow{u}, \overrightarrow{\mathcal{Y}}{=}\overrightarrow{y}]V{=}v$ for fixed $\overrightarrow{u}$, $\overrightarrow{y}$ and $V$. Then, the set of structural functions for $\mathcal{V}$ in $\Gamma$ defined as $\mathcal{F}^\Gamma := \{f_V^\Gamma \mid V \in \mathcal{V}\}$.*

- *The valuation $\mathcal{A}^\Gamma$ is defined, for every $Z \in \mathcal{U} \cup \mathcal{V}$, as*

$$\mathcal{A}^\Gamma(Z) = z \quad \text{if and only if} \quad Z{=}z \in \Gamma$$

  *Note: axioms HP1 and HP2 ensure that $\mathcal{A}^\Gamma$ is a well-defined function, as they guarantee $\Gamma$ has one and only one formula of the form $Z{=}z$ for a fixed $Z$.*

We show that the structure just defined is indeed a causal model.

**4.** Proposition. *Take $\Gamma \in \mathbb{C}$. The tuple $\langle \mathcal{S}, \mathcal{F}^\Gamma, \mathcal{A}^\Gamma \rangle$ is a proper causal model, that is, **(i)** $\mathcal{F}^\Gamma$ is recursive, and **(ii)** $\mathcal{A}^\Gamma$ complies with $\mathcal{F}^\Gamma$.*

*Proof.*

***(i)*** Suppose $\mathcal{F}^\Gamma$ is not recursive, i.e., suppose $\hookrightarrow^+_{\mathcal{F}^\Gamma}$ is either not asymmetric or else not transitive. The relation is transitive by construction, so the problem should be asymmetry: there are $X_1, X_2 \in \mathcal{U} \cup \mathcal{V}$ such that $X_1 \hookrightarrow^+_{\mathcal{F}^\Gamma} X_2$ and $X_2 \hookrightarrow^+_{\mathcal{F}^\Gamma} X_1$, that is,

$$X_1 \hookrightarrow_{\mathcal{F}^\Gamma} Y_1 \hookrightarrow_{\mathcal{F}^\Gamma} \cdots \hookrightarrow_{\mathcal{F}^\Gamma} Y_p \hookrightarrow_{\mathcal{F}^\Gamma} X_2, \qquad X_2 \hookrightarrow_{\mathcal{F}^\Gamma} W_1 \hookrightarrow_{\mathcal{F}^\Gamma} \cdots \hookrightarrow_{\mathcal{F}^\Gamma} W_q \hookrightarrow_{\mathcal{F}^\Gamma} X_1$$

Now, note how, for any two variables $Z_1, Z_2 \in \mathcal{U} \cup \mathcal{V}$, if $Z_1 \hookrightarrow_{\mathcal{F}^\Gamma} Z_2$ then $Z_1 \rightsquigarrow Z_2 \in \Gamma$.[18] Thus, all formulas in

$$\left\{ \begin{array}{l} X_1 \rightsquigarrow Y_1,\ Y_1 \rightsquigarrow Y_2,\ \ldots, Y_{p-1} \rightsquigarrow Y_p,\ Y_p \rightsquigarrow X_2, \\ X_2 \rightsquigarrow W_1,\ W_1 \rightsquigarrow W_2,\ \ldots, W_{q-1} \rightsquigarrow W_q,\ W_q \rightsquigarrow X_1 \end{array} \right\}$$

are in $\Gamma$, and so is their conjunction. But, by axiom HP6, $(X_1 \rightsquigarrow Y_1 \wedge \cdots \wedge W_{q-1} \rightsquigarrow W_q) \rightarrow \neg(W_q \rightsquigarrow X_1) \in \Gamma$. This makes $\Gamma$ inconsistent; a contradiction.

***(ii)*** Suppose $\mathcal{A}^\Gamma$ does not comply with $\mathcal{F}^\Gamma$. Then, there is $V \in \mathcal{V}$ such that $\mathcal{A}^\Gamma(V) = v$ but $f^\Gamma_V(\mathcal{A}^\Gamma(\overrightarrow{U}), \mathcal{A}^\Gamma(\overrightarrow{Y})) \neq v$, with $\overrightarrow{U}$ the tuple of all exogenous variables and $\overrightarrow{Y}$ the tuple of all endogenous variables in $\mathcal{V} \setminus \{V\}$. Take $\mathcal{A}^\Gamma(\overrightarrow{U}) = \overrightarrow{u}$ and $\mathcal{A}^\Gamma(\overrightarrow{Y}) = \overrightarrow{y}$.

From $\mathcal{A}^\Gamma$'s definition, $\mathcal{A}^\Gamma(\overrightarrow{U}) = \overrightarrow{u}$, $\mathcal{A}^\Gamma(\overrightarrow{Y}) = \overrightarrow{y}$ and $\mathcal{A}^\Gamma(V) = v$ imply that the formulas in $\{V{=}v\} \cup \{U_i{=}u_i \mid U_i \in \overrightarrow{U}\} \cup \{Y_i{=}y_i \mid Y_i \in \overrightarrow{Y}\}$ are all in $\Gamma$. This and axiom HP3 imply that $[\overrightarrow{U}{=}\overrightarrow{u}, \overrightarrow{Y}{=}\overrightarrow{y}]V{=}v \in \Gamma$. But, from $f^\Gamma_V$'s definition, $f^\Gamma_V(\mathcal{A}^\Gamma(\overrightarrow{U}), \mathcal{A}^\Gamma(\overrightarrow{Y})) \neq v$ implies $[\overrightarrow{U}{=}\overrightarrow{u}, \overrightarrow{Y}{=}\overrightarrow{y}]V{=}v \notin \Gamma$, a contradiction.

We have so far been using expressions of the form $\overrightarrow{X} = \overrightarrow{x}$ ("assignments") only inside intervention modalities. From this point onwards we follow the literature and we allow such expressions to occur also outside of modalities; in such contexts, they must be understood as *conjunctions* of atoms, such as $X_1 = x_1 \wedge \cdots \wedge X_n = x_n$.

**1.** Lemma (Inverse of composition). *Let $\overrightarrow{X}, \overrightarrow{Y}, \overrightarrow{Z}$ be tuples of variables in $\mathcal{U} \cup \mathcal{V}$, and $\overrightarrow{x}\,\overrightarrow{y}\,z \in \mathcal{R}(\overrightarrow{X}\,\overrightarrow{Y}\,\overrightarrow{Z})$. From the assumptions $[\overrightarrow{X}{=}\overrightarrow{x}]\overrightarrow{Y}{=}\overrightarrow{y}$ and $[\overrightarrow{X}{=}\overrightarrow{x}, \overrightarrow{Y}{=}\overrightarrow{y}]\overrightarrow{Z} = \overrightarrow{z}$ we can formally prove $[\overrightarrow{X}{=}\overrightarrow{x}]\overrightarrow{Z}{=}\overrightarrow{z}$ in $\mathsf{L}_{KC}$.*

*Proof.* Suppose for the sake of contradiction that the set $\Delta = \{[\overrightarrow{X} = \overrightarrow{x}]\overrightarrow{Y} = \overrightarrow{y}, [\overrightarrow{X} = \overrightarrow{x}, \overrightarrow{Y} = \overrightarrow{y}]\overrightarrow{Z} = \overrightarrow{z}\ , \neg[\overrightarrow{X} = \overrightarrow{x}]\overrightarrow{Z} = \overrightarrow{z}\}$ is consistent. If $|(\mathcal{R}(\overrightarrow{Z})| = 1$, this contradicts axiom HP2; so assume $|\mathcal{R}(\overrightarrow{Z})| > 1$. By applying RH2, HP2 and classical logic to the last of these formulas, we obtain that also $\Delta' = \{[\overrightarrow{X} =$

[18] Indeed, let $\overrightarrow{Z^-}$ be a vector containing all variables in $(\mathcal{U} \cup \mathcal{V}) \setminus \{Z_1, Z_2\}$, and suppose $Z_1 \hookrightarrow_{\mathcal{F}^\Gamma} Z_2$. By definition of $\hookrightarrow_{\mathcal{F}^\Gamma}$, there is a vector $\overrightarrow{z^-} \in \mathcal{R}(\overrightarrow{Z^-})$ and there are $z_1, z_1' \in \mathcal{R}(Z_1)$ with $z_1 \neq z_1'$ such that, if $f^\Gamma_{Z_2}(\overrightarrow{z^-}, z_1) = z_2$ and $f^\Gamma_{Z_2}(\overrightarrow{z^-}, z_1') = z_2'$ (with $f^\Gamma_{Z_2}$ the structural function for $X_2$ in $\mathcal{F}^\Gamma$), then $z_2 \neq z_2'$. Thus, from the definition of the structural functions in $\mathcal{F}^\Gamma$, it follows that $[\overrightarrow{Z^-}{=}\overrightarrow{z^-}, Z_1{=}z_1]Z_2{=}z_2 \in \Gamma$ and $[\overrightarrow{Z^-}{=}\overrightarrow{z^-}, Z_1{=}z_1']Z_2{=}z_2' \in \Gamma$ for $\overrightarrow{z^-} \in \mathcal{R}(\overrightarrow{Z^-})$, $z_1 \neq z_1'$ and $z_2 \neq z_2'$. Since $\Gamma$ is maximally consistent, the conjunction of both formulas is also in $\Gamma$, and hence so is $Z_1 \rightsquigarrow Z_2$.

$\overrightarrow{x}]\overrightarrow{Y} = \overrightarrow{y}, [\overrightarrow{X} = \overrightarrow{x}, \overrightarrow{Y} = \overrightarrow{y}]\overrightarrow{Z} = \overrightarrow{z}\ , [\overrightarrow{X} = \overrightarrow{x}]\overrightarrow{Z} = \overrightarrow{z}'\}$ is consistent, for some $\overrightarrow{z}' \neq \overrightarrow{z}$. Applying HP3 to the first and third formulas of $\Delta'$, we obtain $[\overrightarrow{X} = \overrightarrow{x}, \overrightarrow{Y} = \overrightarrow{y}]\overrightarrow{Z} = \overrightarrow{z}'$; by HP1 we obtain $\neg[\overrightarrow{X} = \overrightarrow{x}, \overrightarrow{Y} = \overrightarrow{y}]\overrightarrow{Z} = \overrightarrow{z}$, contradicting the consistency of $\Delta'$.

The following proposition is the crucial part of the proof: it shows that $\langle \mathcal{S}, \mathcal{F}^\Gamma, \mathcal{A}^\Gamma \rangle$ satisfies all 'atoms' (formulas of the form $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$) in $\Gamma$.

**5.** Proposition. *Let $\Gamma \in \mathbb{C}$ be a maximally $\mathsf{L}_{KC}$-consistent set of $\mathcal{L}_{KC}$-formulas. Let $\overrightarrow{X}{=}\overrightarrow{x}$ be an assignment, for $\overrightarrow{X}$ a tuple of variables in $\mathcal{U} \cup \mathcal{V}$; take $Z \in \mathcal{U} \cup \mathcal{V}$ and $z \in \mathcal{R}(Z)$. Then,*

$$[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \in \Gamma \qquad \textit{if and only if} \qquad \langle \mathcal{S}, \mathcal{F}^\Gamma, \mathcal{A}^\Gamma \rangle \models [\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$$

*Proof.* From the semantic interpretation, the right-hand side $\langle \mathcal{S}, \mathcal{F}^\Gamma, \mathcal{A}^\Gamma \rangle \models [\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$ is equivalent to ${\mathcal{A}^\Gamma}^{\mathcal{F}^\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}(Z) = z$. Then, the proof will show that, for any assignment $\overrightarrow{X}{=}\overrightarrow{x}$ on $\mathcal{U} \cup \mathcal{V}$, any $Z \in \mathcal{U} \cup \mathcal{V}$ and any $z \in \mathcal{R}(Z)$,

$$[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \in \Gamma \qquad \text{if and only if} \qquad {\mathcal{A}^\Gamma}^{\mathcal{F}^\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}(Z) = z$$

There are two main cases. First, suppose $Z \in \mathcal{U}$, and take any $\overrightarrow{X}{=}\overrightarrow{x}$.

- Suppose further that $Z$ occurs in $\overrightarrow{X}$, so $Z = X_k$ for some $1 \leqslant k \leqslant |\overrightarrow{X}|$. **(⇒)** Suppose $[\overrightarrow{X}{=}\overrightarrow{x}]X_k{=}z \in \Gamma$. By axiom HP4, we also have $[\overrightarrow{X}{=}\overrightarrow{x}]X_k{=}x_k \in \Gamma$; thus, axiom HP1 and the consistency of $\Gamma$ imply $z = x_k$. Now, from the definition of the value of intervened variables after an intervention (Definition 5), it follows that ${\mathcal{A}^\Gamma}^{\mathcal{F}^\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}(X_k) = x_k$; this, together with $z = x_k$, produces the required ${\mathcal{A}^\Gamma}^{\mathcal{F}^\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}(X_k) = z$. **(⇐)** Suppose ${\mathcal{A}^\Gamma}^{\mathcal{F}^\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}(X_k) = z$. From Definition 5 again, ${\mathcal{A}^\Gamma}^{\mathcal{F}^\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}(X_k) = x_k$, so $z = x_k$. Now, by axiom HP4 again, $[\overrightarrow{X}{=}\overrightarrow{x}]X_k{=}x_k \in \Gamma$ so, since $z = x_k$, it follows that $[\overrightarrow{X}{=}\overrightarrow{x}]X_k{=}z \in \Gamma$.
- Suppose $Z$ does not occur in $\overrightarrow{X}$. By axiom EX, $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \in \Gamma$ if and only if $Z{=}z \in \Gamma$; by the definition of $\mathcal{A}^\Gamma$ (Definition 13), $Z{=}z \in \Gamma$ if and only if $\mathcal{A}^\Gamma(Z) = z$; by the definition of the value an intervened valuation assigns to a non-intervened exogenous variable (Definition 5), $\mathcal{A}^\Gamma(Z) = z$ if and only if ${\mathcal{A}^\Gamma}^{\mathcal{F}^\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}(Z) = z$.

Suppose now $Z \in \mathcal{V}$. The proof proceeds by induction on the number of *non-intervened endogenous* variables, i.e., by induction on the size of $\mathcal{V} \setminus \overrightarrow{X}$.

**Case $|\mathcal{V} \setminus \overrightarrow{X}| = 0$.** This is the case when every endogenous variable is being intervened; in particular, $Z$ is. Then, the argument for the case $Z \in \mathcal{U}$ with $Z$ occurring in $\overrightarrow{X}$ shows that the equivalence holds.

**Case $|\mathcal{V} \setminus \vec{X}| = 1$.** If $Z$ is being intervened (i.e., $Z$ occurs in $\vec{X}$), then the argument for the case $|\mathcal{V} \setminus \vec{X}| = 0$ is enough.

If $Z$ is the lone non-intervened endogenous variable, $\vec{X}$ contains all variables in $\mathcal{V} \setminus \{Z\}$. Then, define $\vec{U'}{=}\vec{u'}$ as the assignment over the exogenous variables not in $\vec{X}$ (i.e., $U' \in \vec{U'}$ if and only if both $U' \in \boldsymbol{\mathcal{U}}$ and $U' \notin \vec{X}$) by taking $u'_i := \mathcal{A}^\Gamma(U'_i)$. From the definition of $\mathcal{A}^\Gamma$, it is clear that $U'_i{=}u'_i \in \Gamma$ for all $U'_i \in \vec{U'}$. Note how the disjoint vectors $\vec{X}$ and $\vec{U'}$ contain, together, exactly all the variables in $(\boldsymbol{\mathcal{U}} \cup \mathcal{V}) \setminus \{Z\}$.Notice that, by the definition of intervention (Definition 5), we have $\mathcal{A}^{\Gamma^{\mathcal{F}^\Gamma_{\vec{X}=\vec{x}}}}(Z) = \mathcal{A}^{\Gamma^{\mathcal{F}^\Gamma_{\vec{X}\vec{U'}=\vec{x}\vec{u'}}}}(Z) = f^\Gamma_Z(\vec{x}, \vec{u}')$. But then, by the construction of $f^\Gamma_Z$ (Definition 13) we have $\mathcal{A}^{\Gamma^{\mathcal{F}^\Gamma_{\vec{X}=\vec{x}}}}(Z) = z$ if and only if $[\vec{X}{=}\vec{x}, \vec{U'}{=}\vec{u'}]Z{=}z \in \Gamma$. In the presence of $[\vec{X}{=}\vec{x}]\vec{U'}{=}\vec{u'} \in \Gamma$ (a consequence of the previous $\vec{U'}{=}\vec{u}' \in \Gamma$ and axiom EX), the latter is equivalent to the required $[\vec{X} = \vec{x}]Z = z$ (by Lemma 1 in one direction, and by axiom HP3 in the other).

**Case $|\mathcal{V} \setminus \vec{X}| = k > 1$.** If $Z$ is being intervened, equivalence follows as shown in the case $|\mathcal{V} \setminus \vec{X}| = 0$.

Suppose $Z$ is not being intervened. Define $\vec{U'}{=}\vec{u'}$ as in the previous case.

**(⇒)** Suppose $[\vec{X}{=}\vec{x}]Z{=}z \in \Gamma$. Based on this, we will build a complete valuation $\mathcal{A}^*$, and we will show that $\mathcal{A}^*$ ***(i)*** agrees with $\mathcal{A}^\Gamma$ on the values of all exogenous variables not in $\vec{X}$, ***(ii)*** follows $\vec{X}{=}\vec{x}$ for the values of exogenous variables in $\vec{X}$, and ***(iii)*** complies with all structural functions in $\mathcal{F}^\Gamma_{\vec{X}=\vec{x}}$. Since there is a unique valuation satisfying these three requirements ($\mathcal{F}^\Gamma$ is recursive, as shown in Proposition 4), it will follow that $\mathcal{A}^* = \mathcal{A}^{\Gamma^{\mathcal{F}^\Gamma_{\vec{X}=\vec{x}}}}$. As it will be shown, $\mathcal{A}^*(Z) = z$, so that will produce the required $\mathcal{A}^{\Gamma^{\mathcal{F}^\Gamma_{\vec{X}=\vec{x}}}}(Z) = z$.

Recall that $\vec{U'}$ contains exactly all exogenous variables not in $\vec{X}$; let $\vec{V'}$ be the vector containing exactly all *endogenous* variables not in $\vec{X}$. Then, define

- $\mathcal{A}^*(X_i) := x_i$ for $X_i \in \vec{X}$;
- $\mathcal{A}^*(U'_i) := u'_i$ for $U'_i \in \vec{U'}$;
- $\mathcal{A}^*(V'_i) := v'_i$ if and only if $[\vec{X}{=}\vec{x}]V'_i{=}v'_i \in \Gamma$, for $V'_i \in \vec{V'}$.[19]

Note how ***(i)*** $\mathcal{A}^*$ agrees with $\mathcal{A}^\Gamma$ on the values of all exogenous variables not in $\vec{X}$ (i.e., variables in $\vec{U'}$) because $\vec{u'}$ is directly taken from $\mathcal{A}^\Gamma$. Moreover, ***(ii)*** it follows $\vec{X}{=}\vec{x}$ for the values of all (in particular, the exogenous) variables in

[19] Axioms HP1 and HP2 guarantee that this uniquely determines the value of each variable in $\vec{V'}$.

$\overrightarrow{X}$. Then, ***(iii)*** it is only left to show that $\mathcal{A}^*$ complies with $\mathcal{F}^{\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}$. For notation, use $y^*$ to denote the value a variable $Y$ receives according to $\mathcal{A}^*$. Note how, since $|\mathcal{V} \setminus \overrightarrow{X}| > 1$, there are at least 2 endogenous variables that are not being intervened (i.e., there are at least two variables in $\overrightarrow{V'}$); denote them by $W_1$ and $W_2$. By definition of the values in $\overrightarrow{v'}$, we have $[\overrightarrow{X}{=}\overrightarrow{x}]W_1{=}w_1^* \in \Gamma$ and $[\overrightarrow{X}{=}\overrightarrow{x}]W_2{=}w_2^* \in \Gamma$.

For the proof, it should be shown that, for every *endogenous* variable $Y$, the value $y^*$ complies with the structural function for $Y$ in $\mathcal{F}^{\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}$. Take any endogenous variable $Y$ different from $W_1$. If $Y$ is in $\overrightarrow{X}$, from axiom HP4 it follows that $[\overrightarrow{X}{=}\overrightarrow{x}, W_1{=}w_1^*]Y{=}y^* \in \Gamma$. Otherwise, $Y$ is not in $\overrightarrow{X}$, so $Y$ is in $\overrightarrow{V'}$ and therefore $[\overrightarrow{X}{=}\overrightarrow{x}]Y{=}y^* \in \Gamma$. But $[\overrightarrow{X}{=}\overrightarrow{x}]W_1{=}w_1^* \in \Gamma$ so, by axiom HP3, $[\overrightarrow{X}{=}\overrightarrow{x}, W_1{=}w_1^*]Y{=}y^* \in \Gamma$. Thus, $[\overrightarrow{X}{=}\overrightarrow{x}, W_1{=}w_1^*]Y{=}y^* \in \Gamma$ holds for every $Y \in \mathcal{V}$ different from $W_1$. Since $|\mathcal{V} \setminus (\overrightarrow{X} \cup \{W_1\})| = k - 1$, from inductive hypothesis it follows that $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x},W_1=w_1^*}(Y) = y^*$, and also that $\mathcal{A}^*$ complies with the structural function for $Y$ from $\mathcal{F}^{\Gamma}_{\overrightarrow{X}=\overrightarrow{x},W_1=w_1^*}$, since $\mathcal{A}^*$ agrees with $\mathcal{A}^{\Gamma}$ outside of $\{\overrightarrow{X}, W_1\}$. But $Y$ is different from $W_1$, so $\mathcal{A}^*$ complies with the structural function for $Y$ from $\mathcal{F}^{\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}$.

Thus, for any $Y$ different from $W_1$, the valuation $\mathcal{A}^*$ complies with the structural function for $Y$ at $\mathcal{F}^{\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}$. An analogous reasoning shows that, for any $Y$ different from $W_2$, the valuation $\mathcal{A}^*$ complies with the structural function for $Y$ at $\mathcal{F}^{\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}$. Thus, for every endogenous variable $Y$, the valuation $\mathcal{A}^*$ complies with the structural function for $Y$ at $\mathcal{F}^{\Gamma}_{\overrightarrow{X}=\overrightarrow{x}}$. This proves ***(iii)***, so we get the desired $\mathcal{A}^* = \mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x}}$. For the final detail, note how our variable $Z$ is in $\overrightarrow{V'}$; since we have assumed $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \in \Gamma$, we have $\mathcal{A}^*(Z) = z$, that is, $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x}}(Z) = z$, as required.

**(⇐)** Suppose $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x}}(Z) = z$. Since $|\mathcal{V} \setminus \overrightarrow{X}| = k > 1$, there are at least two endogenous variables not in $\overrightarrow{X}$. One of them is $Z$; let $W$ be one of the others, and let $w \in \mathcal{R}(W)$ be the value satisfying $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x}}(W) = w$.

- Consider the valuation $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x},W=w}$. Since $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x}}$ and $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x},W=w}$ agree on $W$, it follows that $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x},W=w}(Z) = z$. As $|\mathcal{V} \setminus (\overrightarrow{X} \cup \{W\})| = k - 1$, from the inductive hypothesis it follows that $[\overrightarrow{X}{=}\overrightarrow{x}, W{=}w]Z{=}z \in \Gamma$.
- Consider the valuation $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x},Z=z}$. Since $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x}}$ and $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x},Z=z}$ agree on $Z$, it follows that $\mathcal{A}^{\Gamma\mathcal{F}^{\Gamma}}_{\overrightarrow{X}=\overrightarrow{x},Z=z}(W) = w$. As $|\mathcal{V} \setminus (\overrightarrow{X} \cup \{Z\})| = k - 1$, from the inductive

hypothesis it follows that $[\overrightarrow{X}{=}\overrightarrow{x}, Z{=}z]W{=}w \in \Gamma$.

Thus, $[\overrightarrow{X}{=}\overrightarrow{x}, W{=}w]Z{=}z \in \Gamma$ and $[\overrightarrow{X}{=}\overrightarrow{x}, Z{=}z]W{=}w \in \Gamma$. Then, by axiom HP5, $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \in \Gamma$, as required.

Having proved this 'truth Lemma' for 'atoms' in $\mathcal{L}_{KC}$, the next step is to go from the causal model $\langle \mathcal{S}, \mathcal{F}^{\Gamma}, \mathcal{A}^{\Gamma} \rangle$ to an epistemic causal model where all formulas in $\Gamma$ are satisfied. The definition and lemma are below.

**14.** Definition. *Take* $\Gamma \in \mathbb{C}$.

- *Let* $\mathbb{D}^{\Gamma} := \{\Gamma' \in \mathbb{C} \mid \mathcal{F}^{\Gamma'} = \mathcal{F}^{\Gamma}\}$ *be the set maximally consistent sets in* $\mathbb{C}$ *whose structural functions coincide with those of* $\Gamma$. *Obviously,* $\Gamma \in \mathbb{D}^{\Gamma}$.
- *Define* $R^{\Gamma} \subseteq \mathbb{D}^{\Gamma} \times \mathbb{D}^{\Gamma}$ *as* $(\Gamma_1, \Gamma_2) \in R^{\Gamma}$ *if and only if* $K\chi \in \Gamma_1$ *implies* $\chi \in \Gamma_2$ *for every* $\chi \in \mathcal{L}_{KC}$. *This is the standard definition of the relation in modal canonical models (see, e.g., Fagin et al. [1995], Blackburn et al. [2001]). The elements of* $\mathbb{D}$ *are maximally* $\mathsf{L}_{KC}$*-consistent sets, and* $\mathsf{L}_{KC}$ *includes axioms T, 4 and 5; thus, it follows from standard modal results (see, e.g., the just mentioned reference) that* $R^{\Gamma}$ *is an equivalence relation. In particular, axiom T implies* $(\Gamma, \Gamma) \in R^{\Gamma}$.
- *Define* $\mathcal{T}^{\Gamma} := \{\mathcal{A}^{\Gamma'} \mid (\Gamma, \Gamma') \in R^{\Gamma}\}$ *as containing the valuation function (see Definition 13) of each maximally consistent set in* $\mathbb{D}$ *that is* $R^{\Gamma}$*-reachable from* $\Gamma$. *In particular, from* $(\Gamma, \Gamma) \in R^{\Gamma}$ *it follows that* $\mathcal{A}^{\Gamma} \in \mathcal{T}^{\Gamma}$.

*The structure* $\mathrm{E}^{\Gamma}$ *is given by* $\langle \mathcal{S}, \mathcal{F}^{\Gamma}, \mathcal{T}^{\Gamma} \rangle$.

**2.** Lemma (Truth lemma for $\mathcal{L}_{KC}$). *Take* $\Gamma \in \mathbb{C}$; *recall that* $\mathcal{A}^{\Gamma} \in \mathcal{T}^{\Gamma}$. *Then,*

$$(\langle \mathcal{S}, \mathcal{F}^{\Gamma}, \mathcal{T}^{\Gamma} \rangle, \mathcal{A}^{\Gamma}) \models \chi \quad \textit{if and only if} \quad \chi \in \Gamma$$

*Proof.* The proof is by induction on $\chi \in \Gamma$.

**Case $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$.** The truth-value of an 'atom' $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$ at $(\langle \mathcal{S}, \mathcal{F}^{\Gamma}, \mathcal{T}^{\Gamma} \rangle, \mathcal{A}^{\Gamma})$ is independent from $\mathcal{T}^{\Gamma}$; then,

$$(\langle \mathcal{S}, \mathcal{F}^{\Gamma}, \mathcal{T}^{\Gamma} \rangle, \mathcal{A}^{\Gamma}) \models [\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \quad \text{if and only if} \quad \langle \mathcal{S}, \mathcal{F}^{\Gamma}, \mathcal{A}^{\Gamma} \rangle \models [\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z$$

By Proposition 5, the right-hand side is equivalent to $[\overrightarrow{X}{=}\overrightarrow{x}]Z{=}z \in \Gamma$.

**Case $\neg\chi$.** Immediate from the inductive hypothesis and the properties of a maximally consistent set.

**Case $\chi_1 \wedge \chi_2$.** Immediate from the inductive hypotheses and the properties of a maximally consistent set.

**Case $K\chi$.** As in the same case in the completeness proof of basic modal logic with respect to relational models (see, e.g., [Fagin et al., 1995, Chapter 3]), using the fact that $\mathsf{L}_{KC}$ contains axiom K and rule N.

It is only left to check that $\langle \mathcal{S}, \mathcal{F}^\Gamma, \mathcal{T}^\Gamma \rangle$ is indeed an epistemic causal model.

**6.** Proposition. *Take* $\Gamma \in \mathbb{C}$. *The tuple* $\langle \mathcal{S}, \mathcal{F}^\Gamma, \mathcal{T}^\Gamma \rangle$ *is such that every valuation in* $\mathcal{T}^\Gamma$ *complies with* $\mathcal{F}^\Gamma$.

*Proof.* Take any $\mathcal{A}^{\Gamma'} \in \mathcal{T}^\Gamma$. Note how $\mathcal{A}^{\Gamma'}$ complies with $\mathcal{F}^{\Gamma'}$ (second item in Proposition 4). But $\mathcal{A}^{\Gamma'} \in \mathcal{T}^\Gamma$, so $(\Gamma, \Gamma') \in R^\Gamma$ and hence $\Gamma' \in \mathbb{D}$, which implies $\mathcal{F}^{\Gamma'} = \mathcal{F}^\Gamma$. Thus, $\mathcal{A}^{\Gamma'}$ complies with $\mathcal{F}^\Gamma$.

Here is, then, the full argument for the strong completeness of $\mathsf{L}_{KC}$ for $\mathcal{L}_{KC}$ in epistemic causal models. Let $\Gamma^-$ be any $\mathsf{L}_{KC}$-consistent set of $\mathcal{L}_{KC}$-formulas. From the enumerability of $\mathcal{L}_{KC}$, the set $\Gamma^-$ can be expanded into a maximally $\mathsf{L}_{KC}$-consistent set $\Gamma$. By Lemma 2, all formulas in $\Gamma^-$ are satisfiable in $(\langle \mathcal{S}, \mathcal{F}^\Gamma, \mathcal{T}^\Gamma \rangle, \mathcal{A}^\Gamma)$, which by Proposition 6 is an epistemic causal model.

### 2.8.2 Syntax of $\mathcal{L}_{COD}$

The signatures used in Barbero and Sandu [2019] are pairs of the form $\langle Dom, Ran \rangle$, where $Dom$ is a set of variables (*not* encoding the distinction between exogenous and endogenous variables) and $Ran$ is defined analogously as the $\mathcal{R}()$ used in this paper. For any fixed such signature $\mathcal{S}$, the language $\mathcal{L}_{COD}$ is defined as

$$\alpha ::= Z{=}z \mid Z \neq z \mid \alpha \wedge \alpha \mid \alpha \vee \alpha \mid \alpha \supset \alpha \mid \vec{X}{=}\vec{x} \boxdot\!\!\rightarrow \alpha$$
$$\phi ::= Z{=}z \mid Z \neq z \mid {=}(\vec{X}; Y) \mid \phi \wedge \phi \mid \phi \vee \phi \mid \alpha \supset \phi \mid \vec{X}{=}\vec{x} \boxdot\!\!\rightarrow \phi$$

for $Z, Y, \vec{X} \in Dom$ and $z \in Ran(Z)$.

Note that the expression $\vec{X} = \vec{x}$ used in the antecedent of counterfactuals is just an abbreviation for a conjunction of the form $X_1 = x_1 \wedge \cdots \wedge X_n = x_n$. One can then have *inconsistent* antecedents, say if $\vec{X} = \vec{x}$ contains conjuncts $X = x$ and $X = x'$ with $x \neq x'$. In such cases the intervention is undefined. The semantic clause given in the main text should be extended so as to have any counterfactual with such an antecedent evaluated as (vacuously) true.

Notice also that the antecedent of the operator $\supset$ (selective implication) is restricted to formulas without occurrences of dependence atoms. The consequents of counterfactuals, instead have no restrictions, and they may contain occurrences of $\boxdot\!\!\rightarrow$.

### 2.8.3 Proof for proposition 1 and 2

As pointed out in the main text, here we show how to translate only the *non-nested* formulas of $\mathcal{L}_{COD}$ into $\mathcal{L}_{PAKC}$. Furthermore, we denote as $\alpha, \beta, \gamma$...non-nested formulas of $\mathcal{L}_{COD}$ that have no occurrences of dependence atoms. We need to define a simple preliminary translation $e$ of such formulas, so that they

may correctly act as public announcements. This will be needed in order to translate formulas of the form $\alpha \supset \psi$.

$$e(Y{=}y) := Y{=}y, \qquad e(\beta \wedge \gamma) := e(\beta) \wedge e(\gamma), \qquad e(\beta \supset \gamma) := e(\beta) \to e(\gamma),$$
$$e(Y{\neq}y) := \neg(Y{=}y), \qquad e(\beta \vee \gamma) := \neg(\neg e(\beta) \wedge \neg e(\gamma)), \qquad e(\vec{X} = \vec{x} \,\Box\!\!\to \gamma) := [\vec{X} = \vec{x}]e(\gamma).$$

We point out two simple properties of the preliminary translation $e$.

**3.** Lemma. *Let $\alpha$ be a non-nested formula of $\mathcal{L}_{COD}$ without occurrences of dependence atoms. Let $\langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle$ be an epistemic causal model. Then, for every $\mathcal{A} \in \mathcal{T}$,*

$$(\langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle, \mathcal{A}) \models e(\alpha) \quad \textit{if and only if} \quad \langle \{\mathcal{A}\}, \mathcal{F} \rangle \models \alpha.$$

*Proof.* A simple induction on the syntax of $\alpha$.

**4.** Lemma. *Let $\alpha$ be a non-nested formula of $\mathcal{L}_{COD}$ without occurrences of dependence atoms. Let $E = \langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle$ be an epistemic causal model. Then $E^{e(\alpha)} = \langle \mathcal{S}, \mathcal{F}, \mathcal{T}^{\alpha} \rangle$.*

*Proof.* By definition, $E^{e(\alpha)}$ differs from $E$ only in that its set of valuations is $\{\mathcal{A} \in \mathcal{T} \mid (E, \mathcal{A}) \models e(\alpha)\}$. But, by Lemma 3, this is equal to $\{\mathcal{A} \in \mathcal{T} \mid \langle \{\mathcal{A}\}, \mathcal{F} \rangle \models \alpha\} = \mathcal{T}^{\alpha}$.

Now we can define the translation of (non-nested formulas of) $\mathcal{L}_{COD}$ into $\mathcal{L}_{PAKC}$.

$$tr(Y{=}y) := K(Y{=}y) \qquad tr(\phi_1 \wedge \phi_2) := tr(\phi_1) \wedge tr_(\phi_2)$$
$$tr(Y{\neq}y) := K(\neg(Y{=}y)) \qquad tr(\vec{X}{=}\vec{x} \,\Box\!\!\to \phi) := [\vec{X}{=}\vec{x}]tr(\phi)$$
$$tr(\alpha \supset \phi) := [e(\alpha)!]tr(\phi)$$
$$tr(\phi \vee \psi) := \bigvee_{S \subseteq A} K\Big([(\bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\phi) \;\wedge\; [(\neg \bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\psi)\Big)$$
$$tr(=(X_1, \ldots, X_n; Y)) := \bigwedge_{\vec{x} \in \mathcal{R}(\vec{X})} \bigvee_{y \in \mathcal{R}(Y)} [(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y{=}y)$$

We need to show for any causal team $\langle \mathcal{T}, \mathcal{F} \rangle$ over a signature $\mathcal{S}$ and any formula $\phi \in \mathcal{L}_{COD}$, we have $\langle \mathcal{T}, \mathcal{F} \rangle \models \phi$ if and only if, for all $\mathcal{A} \in \mathcal{T}$, we have $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle, \mathcal{A}) \models tr(\phi)$. This can be done by induction on the complexity of $\phi$. We write $E$ for $\langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle$.

**Case $Y{=}y$.** $\langle \mathcal{T}, \mathcal{F} \rangle \models Y{=}y$ iff $\mathcal{A}(Y) = y$ for each $\mathcal{A} \in \mathcal{T}$ iff $(E, \mathcal{A}) \models K(Y{=}y)$.

**Case $Y{\neq}y$.** $\langle \mathcal{T}, \mathcal{F} \rangle \models Y{\neq}y$ iff $\mathcal{A}(Y) \neq y$ for each $\mathcal{A} \in \mathcal{T}$ iff $(E, \mathcal{A}) \models K(\neg(Y{\neq}y))$.

**Case $\psi \wedge \chi$.** This case follows immediately from the inductive hypothesis.

**Case $\vec{X}{=}\vec{x} \boxdot\!\!\rightarrow \chi$.** $\langle \mathcal{T}, \mathcal{F}\rangle \models \vec{X}{=}\vec{x} \boxdot\!\!\rightarrow \chi$ iff $\langle \mathcal{T}^{\mathcal{F}}_{\vec{X}=\vec{x}}, \mathcal{F}_{\vec{X}=\vec{x}}\rangle \models \chi$ iff for all $\mathcal{B} \in \mathcal{T}^{\mathcal{F}}_{\vec{X}=\vec{x}}$ we have
$(\langle \mathcal{S}, \mathcal{F}_{\vec{X}=\vec{x}}, \mathcal{T}^{\mathcal{F}}_{\vec{X}=\vec{x}}\rangle, \mathcal{B}) \models tr(\chi)$ iff for all $\mathcal{A} \in \mathcal{T}$ we have $(\langle \mathcal{S}, \mathcal{F}_{\vec{X}=\vec{x}}, \mathcal{T}^{\mathcal{F}}_{\vec{X}=\vec{x}}\rangle, \mathcal{A}_{\vec{X}=\vec{x}}) \models tr(\chi)$ iff for all $\mathcal{A} \in \mathcal{T}$ we have $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}\rangle, \mathcal{A}) \models [\vec{X}{=}\vec{x}]tr(\chi)$.

**Case $\alpha \supset \chi$.** $\langle \mathcal{T}, \mathcal{F}\rangle \models \alpha \supset \chi$ iff $\langle \mathcal{T}^{\alpha}, \mathcal{F}\rangle \models \chi$ iff (by inductive hypothesis) $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}^{\alpha}\rangle, \mathcal{A}) \models tr(\chi)$ for all $\mathcal{A} \in \mathcal{T}^{\alpha}$ iff (by Lemma 4) $(E^{e(\alpha)}, \mathcal{A}) \models tr(\chi)$ for all $\mathcal{A} \in \mathcal{T}^{\alpha}$ iff $(E, \mathcal{A}) \models [e(\alpha)!]tr(\chi)$ for all $\mathcal{A} \in \mathcal{T}^{\alpha}$. For the rest, $(E, \mathcal{A}) \models [e(\alpha)!]tr(\chi)$ holds trivially for every $\mathcal{A} \in \mathcal{T} \setminus \mathcal{T}^{\alpha}$, as $e(\alpha)$ is false on $\mathcal{A}$ by Lemma 3.

**Case $\psi \vee \chi$.** As a preliminary observation, note how the causal team language is *downward closed*, in the sense that if $\langle \mathcal{T}, \mathcal{F}\rangle \models \theta$ and $\mathcal{T}' \subseteq \mathcal{T}$, then $\langle \mathcal{T}', \mathcal{F}\rangle \models \theta$ (see Barbero and Sandu [2020] for a proof). By downward closure, it is easy to see that the statement that there are $\mathcal{T}_1, \mathcal{T}_2$ such that $\mathcal{T}_1 \cup \mathcal{T}_2 = \mathcal{T}$, $\mathcal{T}_1 \models \psi$ and $\mathcal{T}_2 \models \chi$ is equivalent to stating the existence of such $\mathcal{T}_1, \mathcal{T}_2$ which are furthermore disjoint.

Now, write $\vec{Y}$ for $\mathcal{U} \cup \mathcal{V}$; recall that $A$ is the set of all possible assignments to $\vec{Y}$. We have $\langle \mathcal{T}, \mathcal{F}\rangle \models \psi \vee \chi$ iff there are *disjoint* $\mathcal{T}_1 \cup \mathcal{T}_2 = \mathcal{T}$ such that $\langle \mathcal{T}_1, \mathcal{F}\rangle \models \psi$ and $\langle \mathcal{T}_2, \mathcal{F}\rangle \models \chi$ iff

(by inductive hypothesis) there are disjoint $\mathcal{T}_1 \cup \mathcal{T}_2 = \mathcal{T}$ such that $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}_1\rangle, \mathcal{A}) \models tr(\psi)$ for all $\mathcal{A} \in \mathcal{T}_1$ and $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}_2\rangle, \mathcal{A}) \models tr(\chi)$ for all $\mathcal{A} \in \mathcal{T}_2$ iff there are disjoint $\mathcal{T}_1 \cup \mathcal{T}_2 = \mathcal{T}$ such that $(E, \mathcal{A}) \models [(\bigvee_{\mathcal{B} \in \mathcal{T}_1} \vec{Y} = \mathcal{B}(\vec{y}))!]tr(\psi)$ for all $\mathcal{A} \in \mathcal{T}_1$ and $(E, \mathcal{A}) \models [(\bigvee_{\mathcal{B} \in \mathcal{T}_2} \vec{Y} = \mathcal{B}(\vec{y}))!]tr(\chi)$ for all $\mathcal{A} \in \mathcal{T}_2$.

For the next step, notice that the first of these public announcement formulas holds trivially on valuations from $\mathcal{T}_2$ (where the announcement is false); analogously, the second formula holds trivially on valuations from $\mathcal{T}_1$ Thus, the statement above is equivalent to the assertion that both formulas hold on each valuation of $\mathcal{T}$. If furthermore we write $S$ for the set of assignments to $\vec{Y}$ that correspond to valuations in $\mathcal{T}_1$, since $\mathcal{T}_1$ and $\mathcal{T}_2$ are disjoint we can rewrite the statement as: there is an $S \subseteq A$ such that, for all $\mathcal{A} \in \mathcal{T}$, $(E, \mathcal{A}) \models [(\bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\psi) \wedge [(\neg \bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\chi)$.

By the semantic clauses, this is equivalent to saying that the same assertion holds for the same formula preceded by $K$. By classical logic, it follows that we can invert the order of the quantifiers,

$$\begin{aligned} &\text{for all } \mathcal{A} \in \mathcal{T} \text{ there is an } S \subseteq A \text{ such that} \\ &(E, \mathcal{A}) \models K\Big([(\bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\psi) \ \wedge \ [(\neg \bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\chi)\Big). \end{aligned} \tag{$*$}$$

Then this is equivalent to: for all $\mathcal{A} \in \mathcal{T}$, $(E, \mathcal{A}) \models \bigvee_{S \subseteq A} K\Big([(\bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\psi) \wedge [(\neg \bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\chi)\Big)$, i.e. the desired conclusion.

In the opposite direction, assume (*) holds. We need to show that we can swap the two quantifiers; this is not given by a logical rule, but we have instead to show that we can take the same $S$ for all $\mathcal{A}$. But this follows immediately from the clause for $K$: if, for a fixed $S$, we have $(E, \mathcal{A}) \models K\Big([(\bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\psi) \wedge [(\neg \bigvee_{\vec{Y}=\vec{y} \in S} \vec{Y} = \vec{y})!]tr(\chi)\Big)$ for some $\mathcal{A}$, then it holds (with the same $S$) for each $\mathcal{A} \in \mathcal{T}$.

**Case $=(\vec{X}; Y)$.** Let $\vec{X}$ be $X_1, \ldots, X_n$. Suppose $\langle \mathcal{T}, \mathcal{F} \rangle \models =(\vec{X}; Y)$; this holds iff for all $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{T}$, if $\mathcal{A}_1(X_i) = \mathcal{A}_2(X_i)$ for all $1 \leq i \leq n$, then $\mathcal{A}_1(Y) = \mathcal{A}_2(Y)$, that is, iff for all $\vec{x} \in \mathcal{R}(\vec{X})$ there is some $y \in \mathcal{R}(Y)$ such that for all $\mathcal{A} \in \mathcal{T}$, $\mathcal{A}(X_1) = x_1, ..., \mathcal{A}(X_n) = x_n$ implies $\mathcal{A}(Y) = y$, which is equivalent to stating that for all $\vec{x} \in \mathcal{R}(\vec{X})$, there is some $y \in \mathcal{R}(Y)$ such that for all $\mathcal{A} \in \mathcal{T}$, $(E, \mathcal{A}) \models [(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y = y)$. Then, it follows that for all $\mathcal{A} \in \mathcal{T}$, $(E, \mathcal{A}) \models \bigwedge_{\vec{x} \in \mathcal{R}(\vec{X})} \bigvee_{y \in \mathcal{R}(Y)} [(X_1{=}x_1 \wedge ... \wedge X_n{=}x_n)!]K(Y{=}y)$.

In the opposite direction, supposing that for all $\mathcal{A} \in \mathcal{T}$ the above holds, we only need to prove that $y$ can be chosen independently of $\mathcal{A}$ (i.e., only as a function of $x_1, \ldots, x_n$). Actually, we prove that it *must* be chosen independently of $\mathcal{A}$. Suppose for the sake of contradiction that, for some $x_1 \ldots x_n \in \mathcal{R}(\vec{X})$, we have $y \neq y' \in \mathcal{R}(y)$ such that $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle, \mathcal{A}) \models [(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y = y)$ and $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle, \mathcal{A}) \models [(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y{=}y')$. From this we easily get that every assignment $\mathcal{B}$ in the causal epistemic model $\langle \mathcal{S}, \mathcal{F}, \mathcal{T}^{X_1=x_1 \wedge \cdots \wedge X_n=x_n} \rangle$ satisfies both $\mathcal{B}(Y) = y$ and $\mathcal{B}(Y) = y'$, a contradiction.

Similarly, we can prove each of the two claims of Proposition 2 by induction on the complexity of $\phi$. As before, the case for $\wedge$ is trivial. Again, write $E$ for $\langle \mathcal{S}, \mathcal{F}, \mathcal{T} \rangle$. In the case of claim ***(i)***, for all operators except dependence atoms, we can follow word-by-word the left-to-right entailments from the proof of Proposition 1. In the proof of the case $=(\vec{X}; Y)$ we observe, as an additional step, that from the assumption that for all $\vec{x} \in \mathcal{R}(\vec{X})$ there is some $y \in \mathcal{R}(Y)$ such that for all $\mathcal{A} \in \mathcal{T}$, $(E, \mathcal{A}) \models [(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y{=}y)$ we can infer, by the semantic clause for $K$, that the same statement holds for the formula $K[(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y = y)$. It is then immediate to conclude that, for all $\mathcal{A} \in \mathcal{T}$, $(E, \mathcal{A}) \models tr^*(\phi)$.

Let us then prove claim ***(ii)*** of Proposition 2.

**Cases $X{=}x$ and $X{\neq}x$.** Suppose there is $\mathcal{A} \in \mathcal{T}$ such that $(E, \mathcal{A}) \models tr^*(X{=}x)$ (i.e., $K(X{=}x)$). Then, for all $\mathcal{A} \in \mathcal{T}$, $\mathcal{A}(X) = x$, i.e. $\langle \mathcal{T}, \mathcal{F} \rangle \models X{=}x$. The proof for $X{\neq}x$

is analogous.

**Case $\vec{X}{=}\vec{x} \mathrel{\Box\!\!\rightarrow} \chi$.** Suppose $(\langle \mathcal{S}, \mathcal{F}, \mathcal{T}\rangle, \mathcal{A}) \models [\vec{X}{=}\vec{x}]tr^*(\chi)$ holds for some $\mathcal{A} \in \mathcal{T}$. Then, $(\langle \mathcal{S}, \mathcal{F}_{\vec{X}=\vec{x}}, \mathcal{T}^{\mathcal{F}}_{\vec{X}=\vec{x}}\rangle, \mathcal{A}_{\vec{X}=\vec{x}}) \models tr^*(\chi)$ and therefore, by inductive hypothesis, $\langle \mathcal{T}^{\mathcal{F}}_{\vec{X}=\vec{x}}, \mathcal{F}_{\vec{X}=\vec{x}}\rangle \models \chi$, i.e., $\langle \mathcal{T}, \mathcal{F}\rangle \models \vec{X}{=}\vec{x} \mathrel{\Box\!\!\rightarrow} \chi$.

**Case $\alpha \supset \chi$.** Suppose there is $\mathcal{A} \in \mathcal{T}$ such that $(E, \mathcal{A}) \models K[e(\alpha)!]tr(\chi)$. Then for all $\mathcal{A} \in \mathcal{T}$ we have $(E, \mathcal{A}) \models [e(\alpha)!]tr(\chi)$. In particular, this holds for all $\mathcal{A} \in \mathcal{T}^{\alpha} \subseteq \mathcal{T}$, so we can proceed as in the right-to-left direction of the proof of Proposition 1.

**Case $\psi \vee \chi$.** Suppose there is a valuation $\mathcal{A}$ in the set $\mathcal{T}$ satisfying $(E, \mathcal{A}) \models \bigvee_{S\subseteq A} K\Big([(\bigvee_{\vec{Y}=\vec{y}\in S} \vec{Y}{=}\vec{y})!]tr(\phi) \wedge [(\neg \bigvee_{\vec{Y}=\vec{y}\in S} \vec{Y}{=}\vec{y})!]tr(\psi)\Big)$. So there is an $S \subseteq A$ such that, for all $\mathcal{A} \in \mathcal{T}$, $(E, \mathcal{A}) \models [(\bigvee_{\vec{Y}=\vec{y}\in S} \vec{Y} = \vec{y})!]tr(\phi) \wedge [(\neg \bigvee_{\vec{Y}=\vec{y}\in S} \vec{Y} = \vec{y})!]tr(\psi)$. From this point we can proceed as in right-to-left direction of the proof of Proposition 1.

**Case $=(\vec{X}; Y)$.** Suppose there is $\mathcal{A} \in \mathcal{T}$ such that $(E, \mathcal{A}) \models \bigwedge_{\vec{x}\in\mathcal{R}(\vec{X})} \bigvee_{y\in\mathcal{R}(Y)} K[(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y{=}y)$; then, for all $x_1, \ldots, x_n \in \mathcal{R}(X_1, \ldots, X_n)$ there is a $y \in \mathcal{R}(Y)$ such that, for all $\mathcal{A} \in \mathcal{T}$, $(E, \mathcal{A}) \models [(X_1{=}x_1 \wedge \cdots \wedge X_n{=}x_n)!]K(Y{=}y)$. From this point we can proceed as in the right-to-left case of Proposition 1.

# Chapter 3

# A Causal Account of Epistemic Counterfactuals

## 3.1 Introduction

For many years the literature on the meaning of counterfactual conditionals has been dominated by the similarity approach of Stalnaker [Stalnaker, 1968] and Lewis [Lewis, 1973]. According to this approach a counterfactual with antecedent $A$ and consequent $C$ is true in a possible world $w$, if among the worlds that make $A$ true those most similar to $w$ also make $C$ true. The central challenge this approach has to face is to specify the relevant similarity relation correctly.[1] Without proper restrictions on the similarity relation the approach is prone to counterexamples and mispredictions. Recently, we see a wave of causal approaches to counterfactuals, building on the seminal work of Judea Pearl [Pearl, 2009, 2013], cf. Hiddleston [2005], Schulz [2011], Santorio [2019], Ciardelli et al. [2018] among others.[2] These proposals have been able to account for many properties of counterfactual conditionals that have been problematic for the similarity approach. But there is still a group of examples that cannot be treated this way. What they have in common is that they seem to involve epistemic reasoning. Consider, for instance, the following example from Kratzer [1989].

**2.** Example. **(King Ludwig Example)**
*King Ludwig of Bavaria likes to spend his weekends in Leoni Castle. Whenever the Royal Bavarian flag is up and the lights are on, the King is in the Castle. From a distance a traveler observes that the lights are on, the flag is down, and concludes that the King is away. She says ...*

[1] See, for instance Fine [1975] and Lewis [1979] for discussion.

[2] Pearl's approach can to a great extent be understood as working with a very specific similarity relation, see Halpern [2013], Marti and Pinosio [2014a].

(1) If the flag had been up, the King would have been in the castle.

The conditional (1) is generally judged true in the given context. This conditional seems to have a strong epistemic element, which can be brought out with the paraphrase given in (2a), but disappears in (2b). This later counterfactual is intuitively false in this context. Standard causal approaches to counterfactuals can correctly capture our intuitions concerning (2b), but are unable to deal with the acceptability of (1) or its paraphrase (2a), which reason from the observation of a certain effect (the flag being up) to its cause (the king being in the castle).

(2) a. If the flag had been up, I would have thought that the king is in the castle.

b. If the flag had been up, this would have caused the king to be in the castle.

Examples like this are the topic of the present paper. We will develop an approach that can account for them, but still stays within the framework of causal approaches to counterfactuals. In other words, we will not account for the acceptability of (1) by proposing a second epistemic reading for the counterfactual. Instead, we will argue that we need to add epistemic reasoning to the meaning assigned by the causal approach to counterfactuals. More concretely, we propose that counterfactuals reason about what one would have *believed* in case the antecedent had been brought about. In order to formalise this idea we will build on the causal approach to counterfactuals developed in Pearl [2009]. We will extend the structural models used in this approach with a representation of an agent's belief state using Baltag and Smets [2008a]. This allows us to add belief operators to the formal language for causal models introduced in Halpern [2016]. Using this extended language we can then express the proposed meaning of counterfactual conditionals in the object language. This approach still accounts for the examples used to motivate the causal approach to counterfactuals, but is also able to deal with epistemic counterfactuals like (1).

The paper is structured as follows. Section 3.2 introduces the version of the causal approach to counterfactuals that we will work with, which is the formalisation of Pearl's original ideas proposed in Halpern [2016]. In Section 3.3 we will dive in more detail into the problem this approach has with epistemic counterfactuals. We will argue against solving this problem by proposing a second epistemic reading of counterfactuals that is based on belief revision. Our alternative solution will be introduced in the Sections 3.4 and 3.5. In Section 3.4 we will define epistemic causal models plus a formal language talking about these models. Section 3.5 spells out the new proposal for the meaning of counterfactual conditionals and contains a discussion of key examples. In Section 3.6 we will discuss a number of interesting questions this approach gives rise to. Section 3.7 will summarise our results and highlight open questions.

## 3.2 Counterfactuals in terms of interventions

As pointed out in the introduction, in recent years the literature on the meaning of counterfactual conditionals has been dominated by causal approaches. The strength of this line of approach can be nicely illustrated with the following example from Lifschitz.

**3.** Example. **(Circuit Example)**
*Suppose there is a circuit such that the light is on (L) exactly when both switches are in the same position (up or not up). At the moment switch 1 is down ($\neg S_1$), switch 2 is up ($S_2$) and the lamp is out ($\neg L$).*

(3) If switch 1 had been up, the light would have been on.

The challenge that any approach to the meaning of counterfactual conditionals has to face can be described as follows: when we interpret a counterfactual conditional, we consider a hypothetical scenario in which the antecedent is true and check whether the consequent is true as well. But which facts of the actual world do still hold in this hypothetical scenario? In order to account for the truth of (3) in the given context, for instance, the position of the second switch has to be selected as one of the facts that is kept, while the state of the lamp is given up. Otherwise, the truth of the counterfactual (3) cannot be predicted. But why should it be that $S_2$ is fixed, but $L$ is not? A causal approach proposes that the reason is that $S_2$ doesn't causally depend on the variable $S_1$ the antecedent talks about, while $L$ does. In other words, when interpreting counterfactuals, we fix the causally independent facts and vary the facts causally dependent on the antecedent.

Let's have a closer look at how this idea can be made formally precise. For this we will use the formalisation proposed in Halpern [2013, 2016] of the seminal work of Judea Pearl [Pearl, 2009, 2013, Galles and Pearl, 1998b]. Following Pearl counterfactuals should be understood as reasoning about hypothetical *interventions*: given a representation of the relevant causal dependencies, the antecedent is cut loose from the facts it causally depends on and stipulated to hold by law. On the resulting model you run a simulation computing the causal effects of such an intervention and check whether the consequent of the counterfactual becomes true as well. If yes, the counterfactual is predicted to be true. In the circuit example 3 we consider what would have happened, had we manipulated the position of switch 1 to being up. Well, this would have caused the light to be on. Thus, the counterfactual is predicted to be true.

We introduce the formalisation of this idea in three steps. First, we define the notion of a causal model. Causal models are direct representations of causal dependencies between a number of variables. Then, we will introduce a formal language that can talk about these models. This language will contain an

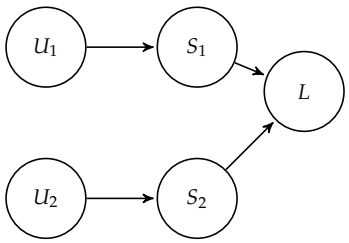
expression of intervention and allow us to formulate counterfactual conditionals. Thus, the approach to the meaning of counterfactuals comes as part of the semantics that will be provided for this formal language.

**Causal models.** A causal model is a pair $(\mathcal{S}, \mathcal{F})$. The *signature* $\mathcal{S}$ fixes a set of causal variables, which represent the objects related by causal dependency. These variables are divided into two kinds: exogenous variables, which do not depend causally on any other variable, and endogenous variables, which do depend on other variables. Formally, a signature is defined as a triple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$ with $\mathcal{U}$ the set of exogenous variables, $\mathcal{V}$ the set of endogenous variables and $\mathcal{R}$ a function that indicates the range of possible values of each causal variable. To model example 3, for instance, we can use three endogenous variables $\mathcal{V} = \{S_1, S_2, L\}$ (already introduced in the example itself), which can each take two values, 1 and 0 (1 stands for *up* and 0 for *down* in case of the switches, and 1 for *on* and 0 for *off* for the lamp). We add two exogenous variables $\mathcal{U} = \{U_1, U_2\}$ that represent factors causally responsible for the position of the switches.

Next, we need to represent the causal dependencies holding between the variables. This is done using structural functions, captured by a function $\mathcal{F}$, which maps each endogenous variable $Y$ to a function $F_Y$ that determines the value of $Y$ given the value of all the other variables in $\mathcal{U} \cup (\mathcal{V} - \{Y\})$.[3] Even though $F_Y$ formally takes into account all other variables, the value of $Y$ might not depend on some of them. In our example, for instance, the values of the switches do not causally depend on the state of the lamp or on each other. We call those variables $X_1, ..., X_n$ that the value of some endogenous variable $Y$ depends on according to $F_Y$ the *causal parents* of $Y$. When we describe $F_Y$ for concrete examples we will only define $F_Y$ for the parents of $Y$. As an example, on the right side of Figure 3.1 we define the functions $F_{S_1}$, $F_{S_2}$ and $F_L$ for our circuit example. As you can see, $U_2$, $S_2$ and $L$, for instance, are not mentioned in the definition of $F_{S_1}$. Using the notion of parents we can also represent the dependencies encoded in a causal model graphically by connecting two variables with an arrow in case the first variable is a causal parent of the second variable. For the circuit example the resulting graph is given on the left side of Figure 3.1. If the graph constructed in this way doesn't contain any loops, the causal model is called recursive. In this paper we will only consider recursive causal models.

**Formal language.** The primitive elements of the language $\mathcal{L}_\mathcal{S}$ for a signature $\mathcal{S}$ are statements claiming that a certain variable takes a particular value: $X = x$, where $X$ is an endogenous variable and $x$ a possible value of this variable.

[3] This model spells out a deterministic conception of causation. However, uncertainty about the values of variables can be introduced in terms of uncertainty about the value of the exogenous variables; in our example $U_1$ and $U_2$.




| $F_{S_1}$: | $U_1$ | $S_1$ |
|---|---|---|
| | 0 | 0 |
| | 1 | 1 |

| $F_{S_2}$: | $U_2$ | $S_2$ |
|---|---|---|
| | 0 | 0 |
| | 1 | 1 |

| $F_L$: | $S_1$ | $S_2$ | $L$ |
|---|---|---|---|
| | 0 | 0 | 1 |
| | 0 | 1 | 0 |
| | 1 | 0 | 0 |
| | 1 | 1 | 1 |

Figure 3.1: A representation of the causal model for the circuit example 3

Based on these primitive elements we build a propositional language in the familiar way. The only new expressions we add are formulas of the form $[Y_1 = y_1, ..., Y_k = y_k]\phi$, where $\phi$ is a Boolean combination of formulas of the form $X = x$, $Y_1, ..., Y_k$ are distinct endogenous variables, and $x \in \mathcal{R}(X)$, $y_1 \in \mathcal{R}(Y_1)$, $y_2 \in \mathcal{R}(Y_2)$, ... . We will often shorten $Y_1 = y_1, ..., Y_k = y_k$ using vector notation and write $\overrightarrow{Y} = \overrightarrow{y}$. The sentences $[\overrightarrow{Y} = \overrightarrow{y}]\phi$ are according to Pearl [2009] and Halpern [2000] the formal counterparts of counterfactual sentences. So, going back to our example 3, the counterfactual in (3) translates as $[S_1 = 1]L = 1$.

**Semantics.** Sentences of the language $\mathcal{L}_\mathcal{S}$ are interpreted with respect to a causal model $M = (\mathcal{S}, \mathcal{F})$ and an assignment of values to the exogenous variables $u$.[4] Notice, that because we restrict ourselves to recursive causal models, given $M$ and $u$ we can compute the value of all endogenous variables based on the functional dependencies encoded in $M$. Let $a^{M,u}$ be the resulting extension of $u$ to all variables $\mathcal{U} \cup \mathcal{V}$ of the model.

An atomic sentence $X = x$ is true given a model $M$ and an assignment function $u$ for the exogenous variables in $M$, $M, u \models X = x$, if $a^{M,u}(X) = x$. Truth for Boolean combinations of atomic formula can be defined in the standard way. The most important part of the semantics is the definition of the truth conditions for counterfactuals $[\overrightarrow{Y} = \overrightarrow{y}]\phi$. The "antecedent" $[\overrightarrow{Y} = \overrightarrow{y}]$ of this sentence is interpreted as performing an intervention on the model: it maps a given causal model to a new model in which the variables in $\overrightarrow{Y}$ are cut off their causal history and forced to the values in $\overrightarrow{y}$. Given a causal model $M = (\mathcal{S}, \mathcal{F})$, we can define the new causal model that results from setting $\overrightarrow{Y}$ to $\overrightarrow{y}$ as $M_{\overrightarrow{Y}=\overrightarrow{y}} = (\mathcal{S}, \mathcal{F}_{\overrightarrow{Y}=\overrightarrow{y}})$ where $\mathcal{F}_{\overrightarrow{Y}=\overrightarrow{y}}$ is the result of replacing the structural functions for $\overrightarrow{Y}$ in $\mathcal{F}$ by $F_{\overrightarrow{Y}}(\overrightarrow{Y}) = \overrightarrow{y}$ (by turning $\mathcal{F}_{Y_1}, ..., \mathcal{F}_{Y_n}$ into constant functions whose output is $y_1, ..., y_n$) and leaving the remaining functions untouched.[5] A formula

[4] Thus, $u$ maps all exogenous variables $U \in \mathcal{U}$ to a value $r \in \mathcal{R}(U)$ this variable can take.

[5] Notice that this operation will again result in a recursive model. Thus, it still holds that

$[\overrightarrow{Y} = \overrightarrow{y}]\phi$ is defined to be true given a causal model $M$ and a setting of the exogenous variables $u$ iff $M_{\overrightarrow{Y}=\overrightarrow{y}}, u \models \phi$. So, this is how counterfactuals are interpreted according to this approach.

Based on the semantics we can now evaluate whether our translation of the counterfactual (3) of Example 3 is true given the causal model $(\mathcal{S}, \mathcal{F})$ we introduced above (see Figure 3.1). Since switch $S_1$ is down and switch $S_2$ is up, the assignment $u$ for the involved exogenous variables maps $U_1$ to 0 and $U_2$ to 1 ($U_1$ and $U_2$ are supposed to be all external factors that determine $S_1$ and $S_2$'s initial state). Intervention with the antecedent $[S_1 = 1]$ brings us to a model $M' = (\mathcal{S}, \mathcal{F}')$, where $\mathcal{F}'_{S_2} = \mathcal{F}_{S_2}$ and $\mathcal{F}'_{L} = \mathcal{F}_{L}$, but $\mathcal{F}'_{S_1} = 1$, i.e. $S_1$ is cut loose from $U_1$ and fixed to the value 1. The extension $a^{M',u}$ of $u$ to all variables given $M'$ maps $S_1$ to 1, $S_2$ to 1 and $L$ to 1. Thus, we predict that the lights are on under the counterfactual assumption that the first switch is up ($M, u \models [S_1 = 1]L = 1$). In other words, the approach can account for the intuition that the counterfactual (3) is true in the context of Example 3.

This is the form of the causal approach to the meaning of counterfactuals that we will work with. This approach has been proven to be very successful and can account for many puzzles concerning the meaning of counterfactuals [Pearl, 2009, 2013, Schulz, 2011, Ciardelli et al., 2018, Santorio, 2019]. We think that the approach can also deal with epistemic counterfactuals when extended in the right way.

## 3.3 What about King Ludwig?

### 3.3.1 King Ludwig with intervention

Let us return to the King Ludwig example from the introduction. We repeat the example here for convenience.

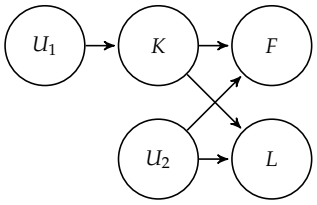
**1.** Example (Revisiting the King Ludwig Example 2). *King Ludwig of Bavaria likes to spend his weekends in Leoni Castle. Whenever the Royal Bavarian flag is up and the lights are on, the King is in the Castle. From a distance a traveler observes that the lights are on, the flag is down, and concludes that the King is away. She says ...*

(1) If the flag had been up, the king would have been in the castle.

A straightforward causal model for this example is given in Figure 3.2. We distinguish three endogenous variables $K$, $F$ and $L$ for the presence of the king, the position of the flag and the state of the lights. All factors that determine King Ludwig's presence are packaged into an exogenous variable $U_1$. According to

given an assignment for the exogenous variables the value of all variables in this model are uniquely defined [Halpern, 2000].

the information provided in Example 2 the position of the flag $F$ and the state of the lamps $L$ are causally dependent on $K$. Since the story leaves it open whether there are other reasons causing the light to be on or the flag to be up in case the king is away, we add an exogenous variable $U_2$ to represent uncertainty about the status of the flag and lights when the king is absent.[6] However, the presence of the king is the only reason for the flag being up and the lamp being on at the same time. This leads to the particular definition of the functional dependencies given in the tables of Figure 3.2.




| $F_K$: | | $F_F$: | | | $F_L$: | | |
|---|---|---|---|---|---|---|---|
| $U_1$ | $K$ | $U_2$ | $K$ | $F$ | $U_2$ | $K$ | $L$ |
| 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | | 2 | 0 | 0 | 2 | 0 | 1 |
| | | 2 | 1 | 1 | 2 | 1 | 1 |
| | | 3 | 0 | 1 | 3 | 0 | 0 |
| | | 3 | 1 | 1 | 3 | 1 | 1 |

Figure 3.2: A causal model for the King Ludwig example

In order to account for the counterfactual (1) we would need its formal translation $[F = 1]K = 1$ to be true with respect to the model above and the setting $U_1 = 0, U_2 = 2$ of the exogenous variables (this is the setting describing the actual world). Obviously, this is not the case. Changing $F$ by intervention will have no (causal) effect on $K$. Thus, the counterfactual (1) is – incorrectly – predicted to be false in the given context. The only thing the causal approach can do is account for the unacceptability of the causal paraphrase of (1) that we gave in (2b) (see p. 40). The problem is that the approach only reasons unidirectionally with the causal flow, while in the King Ludwig example we are asked to reason in the other direction: from effect (flag up) to cause (presence of king).

### 3.3.2 Why not belief revision?

One suggestion often made to overcome the problem mentioned above is that next to the so called *ontic* reading of counterfactuals – which the causal approach captures – we should distinguish a second, epistemic reading of coun-

[6] We could have also used two different exogenous variables for $F$ and $L$, but we try to keep the model as simple as possible.

terfactuals.[7] This second reading should be modelled based on belief revision. However, it is not trivial to make such an approach work for examples like the one at hand. The problem is that we have to explain why when revising the beliefs of the speaker with the proposition that the flag is up, the belief that the lights are on is kept, while the belief that the king is away is given up. One way to solve this issue is to give priority to beliefs that are based on direct observation. In the natural interpretation of the King Ludwig example, for instance, the position of the flag and the status of the lights are directly observed by the speaker, while the fact that the king is away is inferred from these observations. We could propose now that when revising the beliefs of the speaker with the antecedent, we only check consistency with her beliefs based on direct observation (or at least prioritise these beliefs). This approach to belief revision would predict the acceptability of the counterfactual in (1).

While this is a possible strategy to address this and similar epistemic flavored examples, we think it is still worthwhile to look for a solution that can do without multiplying senses of counterfactuals. In fact, there are also some empirical arguments that speak against assuming an ambiguity between a causal (or ontic) reading of counterfactuals and an epistemic one based on belief revision. If we propose that counterfactuals are ambiguous, we also need to be able to explain how this ambiguity is resolved in concrete cases. But that seems to be rather problematic for the particular ambiguity proposed here. Consider, for instance, the following example (Veltman [2005], footnote 21).[8]

**4.** Example. **(Three Sisters)**
*Consider the case of three sisters, Ann, Billie and Carol, who own just one bed, large enough for two of them but too small for all three. Every night at least one of them has to sleep on the floor.*

*Suppose Carol is invisible. Suppose further that you are a proud parent of Ann, and before you go to bed you go in and check the kids. You see that Ann is on the floor, Billie is in bed and Carol (obviously) is also in bed. Now you turn to your spouse and comment:*

*(4) If Ann had been in bed, Carol would have been on the floor.*

Intuitively, the counterfactual (4) seems to be acceptable in the described context. In order to account for this observation we have to explain why when considering the possibility that Ann had been in bed, we keep the fact that Billie is in bed, but are willing to give up the fact that Carol is in bed as well. This example is a variation of the King Ludwig case, but without the causal

[7] Already Lewis suggests that there are different ways to resolve similarity, see, for instance, Lewis [1979].

[8] The text of the example is slightly adapted. The original text refers to another example, which we had to incorporate in our version.

dependency of the latter.[9] The causal approach of Section 3.2 cannot account for the example: Ann's being in bed doesn't *cause* Carol to sleep on the floor. But we could explain the example by proposing that this counterfactual gets an epistemic reading based on the version of belief revision described above: we keep the fact that Billie is in bed, because this is observed by the speaker. The whereabouts of Carol had to be inferred from this observation. But just a slight variation of the context gets us in a situation where the counterfactual (4) becomes unacceptable, however, the epistemic approach would still predict the sentence to hold.

**5.** EXAMPLE. **(Three Sisters with causality)**
*Consider the case of three sisters who own just one bed, large enough for two of them but too small for all three. Every night at least one of them has to sleep on the floor. However, whenever Ann and Carol are both in bed, they will kick off Billie; she will have to sleep on the floor.*

*Suppose Carol is invisible. Suppose further that you are a proud parent of Ann, Billie and Carol. Before you go to bed you go in and check on the kids. As described in the original version, Ann is on the floor, Billie is in bed and Carol (obviously) is also in bed. Now you turn to your spouse and comment:*

(4) If Ann had been in bed, Carol would have been on the floor.

The only change is the addition of the underlined sentence. But now (4) is no longer true: if Ann had been in bed, Billie would have been on the floor, because Ann and Carol would have kicked her off the bed. From an epistemic point of view nothing changed, so it's hard to see how any account using belief revision could predict the counterfactual to be false now, while at the same time account for its truth in the first context. The causal approach can explain the example. In Example 5, and in contrast to Example 4, Billie's sleeping place causally depends on Carol's behaviour. That is the reason why now the position of Carol cannot be freely varied anymore. But why should this example get a causal reading, while the former needs to be read epistemically? The answer cannot simply be: because this accounts for the data.[10] Examples like this make an ambiguity approach hard to defend.

### 3.3.3 Exploring an alternative approach

As mentioned before, in this paper we want to explore the possibility to account for epistemic counterfactuals without multiplying readings of counterfactuals.

[9] Notice that it is implicitly assumed that always two girls will sleep in the bed. Otherwise, the parent wouldn't be able to conclude that Carol must be sleeping in the bed as well.

[10] Notice, that it wouldn't help to claim that the reading changes, because now causality is involved in the context. Then, the King Ludwig example should get a causal reading as well.

More concretely, we will propose that the causal approach to counterfactuals can deal with epistemic counterfactuals, if extended with the possibility of epistemic reasoning. The causal approach is right when claiming that counterfactuals reason about what can be inferred from hypothetically making the antecedent true. But *inferred* in this paraphrase means more than just causally inferred. Epistemic inferences need to be taken into account as well. Another way to put our proposal is this: the consequent of a counterfactual conditional is not a statement about the facts after intervention, but about what the speaker would have believed under these circumstances. For instance, in case of the King Ludwig example, the consequent doesn't reason about what would have *happened* if one intervened in the position of the flag, but about what the speaker would have *believed* in this case. Hoisting the flag would not have brought the king to the castle, but it would have made an observer believe that the king is in the castle. This idea will be worked out in the next two sections in more detail.

## 3.4 Combining causal and epistemic reasoning

### 3.4.1 Causal epistemic models

Before we can make the proposal outlined at the end of the previous section precise, we first need to extend the formal framework of the causal approach introduced in Section 3.2 with an epistemic dimension. This is what will happen in the present section. We will start by adding to a causal model a representation of the epistemic state of an agent. Afterwards, we will enrich the formal language with means to talk about this epistemic state.

**Epistemic models.** A common way to formalise an epistemic state is by using a plausibility ordering over possible worlds (to capture belief) together with an information partition (to capture knowledge).[11] Let us define an epistemic model as a triple $(W, V, \Pi, \trianglelefteq)$ where $W$ is a set of possible worlds and $V$ maps all elements of $W$ to an interpretation of the non-logical vocabulary. $\Pi$ is an information partition over $W$, i.e. for each $w \in W$ $\Pi(w)$ is the set of possible worlds that are epistemically indistinguishable for the agent at $w$. The plausibility ordering $\trianglelefteq$ is a pre-order over $W$ ($w_1 \trianglelefteq w_2$ stands for "$w_1$ is considered to be at least as plausible as $w_2$ by the agent"). We demand that only possible worlds in the same cell of partition $\Pi$ can be related by $\trianglelefteq$; worlds in different cells of the information partition are incomparable with respect to their plausibility. Given such a model, we can say that an agent knows a proposition $\phi$ in $w$ iff $\phi$ holds

[11] See, for instance, Baltag and Smets [2008b].

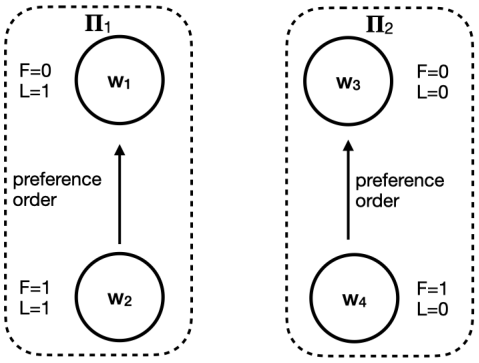



Figure 3.3: An epistemic model for the toy example of Paragraph 3.4.1. The dashed line represents the information partition, the thin arrows stand for the plausibility ordering within each information cell. An arrow points from world $w$ to world $v$ if $v \triangleleft w$. For simplicity, the reflexive loops are not shown in the graph.

in all possible worlds $w' \in \Pi(w)$, and that the agent believes a proposition $\phi$ in $w$ iff $\phi$ holds in the $\trianglelefteq$-minimal worlds of $\Pi(w)$.[12,13]

Let us illustrate these notions using a scenario very similar to the King Ludwig example. Let $F$ stand for the proposition *the flag is up* and $L$ stand for *the light is on*. Based on these two variables we can distinguish four possible worlds: $w_1$ where $L$ holds, but not $F$, $w_2$ where both $F$ and $L$ hold, $w_3$ where neither $F$ nor $L$ hold, $w_4$ where $F$ holds, but not $L$. Furthermore, let us assume that our agent is able to see whether the lights are on, but not whether the flag is up. Then, the agent's knowledge can be captured by the information partition $\Pi(w_1) = \Pi(w_2) = \{w_1, w_2\}; \Pi(w_3) = \Pi(w_4) = \{w_3, w_4\}$. Our agent might, additionally, think that the flag is more likely to be down. That would mean that for this agent $w_1 \triangleleft w_2$ and $w_3 \triangleleft w_4$ (where $w \triangleleft w'$ is shorthand for $w \trianglelefteq w'$ and $w' \not\trianglelefteq w$). A graphical representation of this epistemic model is given in Figure 3.3.

**Merging epistemic models with causal models.** We can combine this notion of an epistemic model with that of a causal model. A causal model doesn't directly distinguish a set of possibilities. But it is very natural to take each combination of possible settings of variables as a possible world and consider the set of all possible combinations of values as the universe. Thus, given a causal model $M = (\mathcal{S}, \mathcal{F})$, we define the universe $W_{\mathcal{S}}$ of this model to be the set of all possible assignment functions from the variables of $\mathcal{S}$ to values these

[12] This formalisation expresses that if an agent is in a possible world $w$, the agent will believe that she is located in the most plausible worlds in $\Pi(w)$.

[13] $w$ is a $\trianglelefteq$-minimal world of $S$ iff there is no $w'$ such that $w' \trianglelefteq w$ and $w \not\trianglelefteq w'$.

variables can take according to $\mathcal{S}$. Notice that this universe will contain worlds that violate causal dependencies encoded in $\mathcal{F}$. This feature of our notion of possible world plays an important role in our approach. With this take on what the set of possibilities is given a causal model, we can now add epistemic structure to this model.

**15.** Definition. *A causal epistemic model is a tuple* $\langle \mathcal{S}, \mathcal{F}, \Pi, \trianglelefteq \rangle$ *that satisfies the following conditions.*

***(i)*** *$\mathcal{S}$ is a triple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$ where $\mathcal{U}$ is the set of exogenous variables and $\mathcal{V}$ is the set of endogenous variables, and $\mathcal{R}$ is a function that indicates the range of possible values of each causal variables.*

***(ii)*** *For each $X \in \mathcal{V}$, $\mathcal{F}_X$ is a function from $(\times_{Z \in \mathcal{U}} \mathcal{R}(Z) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y))$ to $\mathcal{R}(X)$.*

***(iii)*** *Let $W_{\mathcal{S}}$ be the set of possible assignments of values to the variables in $\mathcal{U} \cup \mathcal{V}$. $\Pi$ is a function from $W_{\mathcal{S}}$ to $\wp(W_{\mathcal{S}})$ such that: $w \in \Pi(w)$ for each $w \in W_{\mathcal{S}}$, $w_2 \in \Pi(w_1)$ implies $w_1 \in \Pi(w_2)$, and $w_2 \in \Pi(w_1)$ and $w_3 \in \Pi(w_2)$ implies $w_3 \in \Pi(w_1)$.*

***(iv)*** *$\trianglelefteq$ is a pre-order on $W_{\mathcal{S}}$ that obeys the following constraints: (i) $w_1 \trianglelefteq w_2$ implies $w_1 \in \Pi(w_2)$[14] (ii) for any $w_1, w_2$ such that $w_1 \in \Pi(w_2)$, if there is $X \in \mathcal{V}$ such that $w_1$ complies with $\mathcal{F}_X$ and $w_2$ does not, but there is no $X \in \mathcal{V}$ such that $w_2$ complies $\mathcal{F}_X$ and $w_1$ does not, then $w_1 \triangleleft w_2$ (where $w_1 \triangleleft w_2$ if $w_1 \trianglelefteq w_2$ and $w_2 \not\trianglelefteq w_1$).*

The only non-straightforward part of this definition is condition (*ii*) imposed on the plausibility order. What we express here is the requirement that the agent believes in the causal laws of the model. This is captured by demanding that a possible world that complies with more causal rules is always considered to be more plausible than a possible world in which some of these laws are broken. Thus, we do not *exclude* worlds in which causal laws are broken, but they are considered to be *less plausible* to the agent than other worlds. Going back to the King Ludwig example and the causal model we proposed for this example in Section 3.3.1, a world in which the king is in the castle, the flag is up and the lights are on will always be more plausible for the agent than a world in which the king is in the castle, but the lights, for instance, are off. This is so, because the later world would violate the assumed causal connection between the presence of the king in the castle and the position of the flag/the state of the light.

Notice that condition 4 in Definition 15 formulates constraints for the plausibility order, but doesn't fix the plausibility structure based on the causal laws.

[14] The plausibility ordering is usually assumed to be locally connected in each information cell (i.e. $w_1 \in \Pi(w_2)$ implies $w_1 \trianglelefteq w_2$ or $w_2 \trianglelefteq w_1$). This assumption can be added to the definition proposed here; this change would have no effect on our predictions.

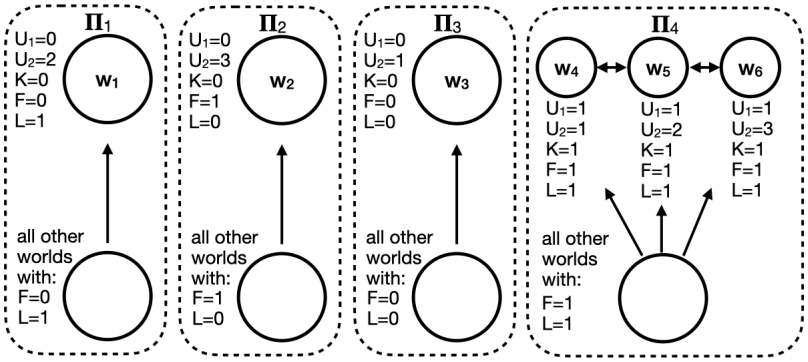



Figure 3.4: The full epistemic model for the King Ludwig Example. The arrows stand for the plausibility ordering (with reflexive loops omitted). Dotted lines represent different cells of the information partition. Within each cell the most plausible worlds are explicitly represented, all other worlds are packed together.

Depending on the context the agent might have other preferences/expectations, which also inform this order. Our definition only claims that, no matter what, the agent will always have a preference for more law-like worlds.

To have a look at a more involved model, let's give the causal epistemic model $\langle \mathcal{S}, \mathcal{F}, \Pi, \trianglelefteq \rangle$ for the King Ludwig example (1). We already provided a causal model for this example, see Figure 3.2 and the discussion in Section 3.3.1. We just need to add to the signature $\mathcal{S}$ and dependency function $\mathcal{F}$ defined there an information partition $\Pi$ and a preference order $\trianglelefteq$ fitting the example.

The universe of the model $W_{\mathcal{S}}$ consists of all possible assignment of values to the causal variables $U_1, U_2, K, F, L$ ($K$ stood for *the king is in the castle*, $F$ for *the flag is up*, and $L$ for *the light is on*, see Section 3.3.1). Since the status of the flag and the light is observable for the agent, she can distinguish between possible worlds whenever the value they assign to $F$ or $L$ differ. In other words for any $w$ and $w' \in W$, $w' \notin \Pi(w)$ if and only if $w$ and $w'$ differ in the value they assign to $F$ or $L$.

Next, let us have a look at the preference order. According to clause 4 of Definition 15, possible worlds that comply with causal laws should always be more plausible than those that do not. It follows that for this example the possible worlds given in the top-row of Figure 3.4 should be preferred over any other (comparable) possible world, as they are the only worlds that obey the causal dependencies encoded in $\mathcal{F}$. Since in the story there is no other information about the plausibility ordering, we can take $\trianglelefteq$ to be the weakest pre-order over $W$ complying with the restriction imposed by the causal laws. The resulting causal epistemic model is sketched in Figure 3.4.

### 3.4.2 A formal language for causal epistemic models

Using the notion of a causal epistemic model we can extend the formal language from Section 3.2 with operators expressing belief (*Bel*) and knowledge (*K*).

**16.** Definition. *Let $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$ be a signature. The extended language for $\mathcal{S}$, $\mathcal{L}^+_{\mathcal{S}}$, is defined as follows:*

$$\alpha ::= V = v \mid \neg\alpha \mid \alpha \wedge \alpha \qquad \text{for } V \in \mathcal{V}, v \in \mathcal{R}(V)$$

$$\phi ::= \alpha \mid \neg\phi \mid \phi \wedge \phi \mid [\overrightarrow{V} = \overrightarrow{v}]\phi \mid Bel\phi \mid K\phi \qquad \text{for } \overrightarrow{V} = (V_1, \ldots, V_n) \in \mathcal{V}^n,\ V_i \neq V_j \text{ for } i \neq j,\ \overrightarrow{v} = (v_1, \ldots, v_n) \text{ with } v_i \in \mathcal{R}(V_i),\ n \in \mathbb{N}$$

The next step is to provide a semantics for this extended language.

**17.** Definition. *Let $M = \langle \mathcal{S}, \mathcal{F}, \Pi, \trianglelefteq \rangle$ be a causal epistemic model and $w \in W_{\mathcal{S}}$ be a possible world for this model, i.e. an assignment for all variables in $\mathcal{S}$. Let $u$ be the restriction of $w$ to the exogenous variables. The truth conditions of the formulas in $\mathcal{L}(\mathcal{S})$ with respect to $M$ and $w$ are defined as follows.*

***(i)*** *For atomic sentences $X = x$, $M, w \models X = x$ iff $w(X) = x$.*

***(ii)*** *The Boolean combinations are defined in the usual way.*

***(iii)*** $M, w \models [\overrightarrow{X} = \overrightarrow{x}]\phi$ *iff* $M, a^{M_{\overrightarrow{X}=\overrightarrow{x}}, u} \models \phi$.[15]

***(iv)*** $M, w \models Bel\phi$ *iff* $M, w' \models \phi$ *for all* $w' \in Min_{\trianglelefteq}(\Pi(w))$

***(v)*** $M, w \models K\phi$ *iff for all* $w' \in \Pi(w)$, $M, w' \models \phi$

The truth conditions for atomic sentences are defined in the obvious way: $M, w \models X = x$ iff the assignment $w$ assigns value $x$ to $X$. For the epistemic operators we apply exactly the same definition of belief and knowledge as in the epistemic logic of Baltag and Smets [2008a]. This leaves us with defining the semantics for sentences of the form $[\overrightarrow{X} = \overrightarrow{x}]\phi$. The definition of their intervention provided here is an extension of the definition given in Section 3.2, but with one subtle difference. In the approach introduced in Section 3.2, formula are evaluated with respect to a model and an assignment of values to the exogenous variables. $[\overrightarrow{X} = \overrightarrow{x}]$ is interpreted as changing the model, and $\phi$ is, then, evaluated with respect to the same assignment for the exogenous variables and the new, manipulated model. In our approach, formula are evaluated with respect to a model and a world: an assignment of values to *all* variables. This gives us the possibility to let intervention move us to a different world, but leave the model intact. This new world is the assignment we get

[15] For the definition of $M_{\overrightarrow{X}=\overrightarrow{x}}$ and $a^{M,u}$ see Section 3.2, p. 43.

by keeping the evaluation of the exogenous variables but recalculating the endogenous variables assuming an intervention forcing $\overrightarrow{X} = \overrightarrow{x}$ in the causal dependencies. We will write $w_{\overrightarrow{X}=\overrightarrow{x}}$ for the resulting assignment whenever the model is clear from the context. $\phi$ is then evaluated with respect to this new possible world $w_{\overrightarrow{X}=\overrightarrow{x}}$ and the *old* model $M$. We want to have the intervention encoded in the world instead of the model, because that is what the plausibility order is looking at. This change allows us to formalise belief in causal laws.[16]

Despite this difference between the two formalisations for the most part the framework of epistemic causal models introduced here is a conservative extension of causal models as defined in Section 3.2. Let us make this a bit more precise. The formal language $\mathcal{L}^+$ introduced here is an extension of the language $\mathcal{L}$ of Section 3.2. Let $\langle \mathcal{S}, \mathcal{F}, \Pi, \trianglelefteq \rangle$ be an epistemic causal model, $u$ an assignment to the exogenous variables of $\mathcal{S}$ and $w = a^{(\mathcal{S},\mathcal{F}),u}$ the extension of $u$ to all variables given $\mathcal{F}$ (see Section 3.2). In this case we can prove that $(\mathcal{S}, \mathcal{F}), u \models \phi$ iff $\langle \mathcal{S}, \mathcal{F}, \Pi, \trianglelefteq \rangle, w \models \phi$ for all sentences $\phi \in \mathcal{L}$, except for iterations of the intervention operator, i.e. formula like $[X = x]([Y = y]\phi)$. So, for the vast majority of formula both approaches give the same results.[17]

## 3.5 A new approach to epistemic counterfactuals

### 3.5.1 Two different ways to reason about interventions

Now that we have our formal framework in place we can come back to our proposal for epistemic counterfactuals. As already stated in Section 3.3.3, we will not propose an ambiguity between two readings of counterfactuals, one based on causation and intervention and one based on belief revision. We propose that counterfactuals are always interpreted as considering a hypothetical intervention. However, we propose that in the resulting hypothetical scenario the agent is not checking whether the consequent holds simpliciter, but whether she would have *believed* the consequent to be true. We can express this difference between our proposal and the standard proposal using the formal language introduced in the previous section. For the counterfactual in (5a) the formula in (5b) describes the interpretation that considers the facts after intervention, the formula in 5c represents the interpretation that considers the beliefs of the agent after intervention. This second formula is the analysis of

[16] Notice that our semantics also works for worlds that violate causal laws. The way the semantics is set up here, these law violations will be ignored for the interpretation of formula $[\overrightarrow{X} = \overrightarrow{x}]\phi$: they get assigned the same truth value as if evaluated in a world with the same interpretation of the exogenous variables, but without any law violations. For now this simple approach taken above is sufficient.

[17] What a correct semantics is for iterated interventions is an interesting, but also complex question. We leave that topic for a different occasion.

counterfactuals that we propose.

(5) a. If $X = x$ had been the case, then $\phi$ would have been true.

b. $[X = x]\phi$

c. $[X = x]Bel\phi$

Figure 3.5 illustrates the difference between the two interpretations in (5b) and (5c). In both cases the antecedent is interpreted as introducing an intervention that moves us to a hypothetical scenario $w_{X=x}$ where the antecedent is true. If the intervention takes place on a variable that is observable, this will also mean that the new world is in a different cell $\Pi(w_{X=x})$ of the information partition. We can now either check directly whether in this hypothetical scenario $w_{X=x}$ the consequent is true – this is expressed by (5b) and results in the causal reading captured by the approach described in Section 3.2. Or we can move to the best world(s) in $\Pi(w_{X=x})$ and check the facts there – this is expressed by (5c) and the analysis that we propose. It is adding epistemic reasoning to the evaluation of counterfactuals, but in a very specific and restricted role.

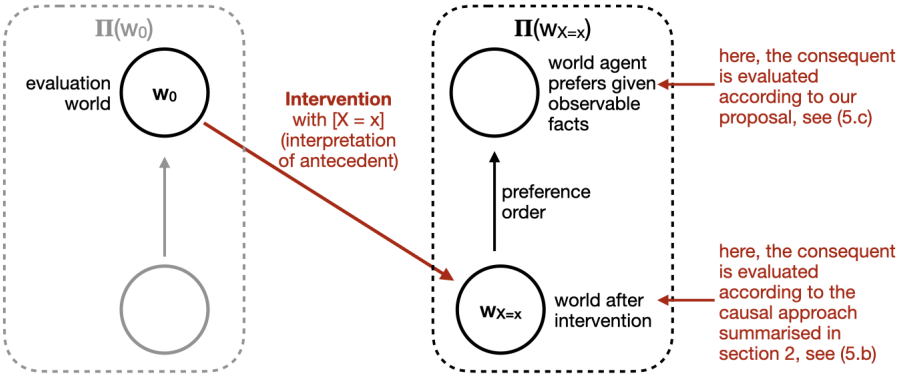



Figure 3.5: Schematic representation of the differences between the truth conditions of (5b) and (5c) for (5a). Only the relevant cells of the information partition are displayed.

For most counterfactuals there is no difference between the truth conditions of (5b) and (5c). Differences can only occur in case the consequent $\phi$ of the conditional contains unobservable variables (otherwise there is no variation in the cell of the partition $\Pi_{X=x}$ that we end up in after intervention). That means that, for instance, in the circuit example 3 discussed in Section 3.2 there is no difference between the two readings and, because no right nested counterfactuals are involved, we make the same predictions as does the approach introduced there. In general, this holds for the gross of the examples discussed in the literature to motivate taking a causal, interventionist approach to counterfactuals. Even if the consequent of a counterfactual contains unobservable

variables, it doesn't automatically mean that (5b) and (5c) will have different truth conditions. If the only thing that counts for the plausibility order are the causal laws, then additionally the variables occurring in the consequent have to be causal ancestors of the variable(s) occurring in the antecedent of the counterfactual for there to be any difference between the two interpretations. The reason is that the only violations of causal laws that might occur in $w_{X=x}$, and thus the only leverage for the plausibility order, are those introduced by the intervention that makes the antecedent true.

### 3.5.2 Accounting for epistemic flavored counterfactuals

With this approach at hand we can now revisit the troublesome examples for counterfactuals discussed in the first part of the paper. Except for the circuit example they all concern situations in which some of the variables are not observable, and thus unknown to the speaker/interpreter (assuming the notion of knowledge formalised here). So, in all these cases our judgements concern our beliefs in the truth of the conditional given what we know and believe about the actual world. Thus, what we need to check in these cases is the truth of the formula $Bel([\overrightarrow{X} = \overrightarrow{x}]Bel\phi)$, where $[\overrightarrow{X} = \overrightarrow{x}]Bel\phi$ is the translation of the relevant counterfactual. In other words, the truth of the counterfactual is checked in the most plausible worlds given what the agent knows. Because the plausibility relation will always take into account the causal laws, in these worlds the facts agree with the laws.[18] For instance, if the agent is in a world in which she can observe some cause, but not its effect, but the laws predict the occurrence of the effect in this case, then she will believe that the effect did occur and evaluate a counterfactual under this assumption. This is illustrated by the grey part (left side) in Figure 3.5. If the context provides us with the information that we are in $\Pi_{w_0}$, then $Bel([\overrightarrow{X} = \overrightarrow{x}]\phi)$ holds if $[\overrightarrow{X} = \overrightarrow{x}]\phi$ is true in the optimal world $w_0$ in this cell of the partition.

**King Ludwig of Bavaria.** Let us start with looking at the predictions this approach makes for the King Ludwig example. We repeat the example a final time.

**2.** Example (**Revisiting the King Ludwig Example**). *King Ludwig of Bavaria likes to spend his weekends in Leoni Castle. Whenever the Royal Bavarian flag is up and the lights are on, the King is in the Castle. From a distance a traveler observes that the lights are on, the flag is down, and concludes that the King is away. She says ...*

(1) If the flag had been up, the king would have been in the castle.

[18] If this is possible given what the agent observes.

We already introduced an epistemic causal model for this example in Section 3.4, Page 51. This was a straightforward extension of the model used in Section 3.2. Just by looking at the model you can already see that this is one of the few cases in which our approach can disagree with the predictions made by the approach introduced in Section 3.2: the consequent talks about a not observable variable that is a causal parent of the variable intervened on in the antecedent. So, let's check whether we predict that $Bel[F = 1]Bel(K = 1)$ holds in the given context.[19] That means that we need to check whether $Bel([F = 1]Bel(K = 1))$ is true at any possible world where $F = 0$ and $L = 1$ hold (see $\Pi_1$ in Figure 3.4). The sentence is a belief statement Hence, in order to be true the formula in scope of the main belief operator has to hold at the most plausible world of the same cell of the partition. This is the world $w$ where the flag is down, the light is on and (now we use the causal laws encoded in the model) the king is not in the castle.[20] $M, w \models [F = 1]Bel(K = 1)$ holds, iff (using Definition 17) $M, w_{F=1} \models Bel(K = 1)$. $w_{F=1}$ is the world we get by setting variable $F$ to value 1 and then recalculating the value of all other endogenous variables based on what $w$ assigns to the exogenous variables and the laws. This will bring us to the world where the flag is up (by intervention), the light is on and the king is away (see Figure 3.6). At this world $w_{F=1}$ the consequent of the conditional, $Bel(K = 1)$, needs to be true. Thus, we have to check whether $K = 1$ holds in all the most plausible worlds in $\Pi(w_{F=1})$. The unique most plausible world in this cell of the partition is $w'$ where the flag is up, the light is on and the king is in the castle. In this world $K = 1$ is obviously true. Thus, $M, w_{F_1} \models Bel(K = 1)$ and, consequently, the sentence $Bel([F = 1]Bel(K = 1))$ holds in the partition where the agent knows that the flag is down and the lights are on. The conditional (1) is predicted to be true by our approach.

Let us shortly reflect on why this approach works. Just as in the causal approach to counterfactuals discussed in Section 3.2, interpreting the antecedent will bring us to a world where the light is on, the flag is up and – the king is away. Intervening in the position of the flag will not bring the king to the castle. However, the agent can only *observe* the variables $F$ and $L$, not the whereabouts of the king. Therefore, in this hypothetical scenario the agent would still *believe* that the king is in the castle. This is why the counterfactual comes out as acceptable.

**Three Sisters – version 2.** In Section 3.6.1 we used two other examples to argue against a proposal that would try to account for our core example by introducing an epistemic reading of counterfactuals based on belief revision. The approach defended here can also account for them. It's worth taking a closer look at these examples, because they involve a non-causal restriction on

[19] To ease notational clutter, we will always write $Bel([F = 1]Bel(K = 1))$ as $Bel[F = 1]Bel(K = 1)$.
[20] Furthermore, the two exogenous variables have the values $U_1 = 0$, $U_2 = 2$.

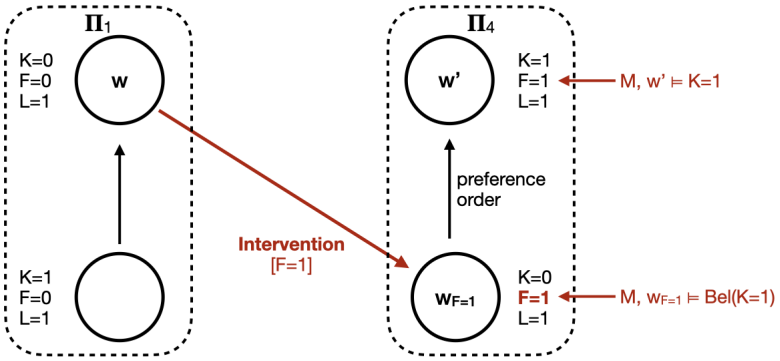



Figure 3.6: A part of the causal epistemic model for the King Ludwig Example. For ease of presentation we collapsed worlds that only differ in the value they assign to $U_1$ and $U_2$.

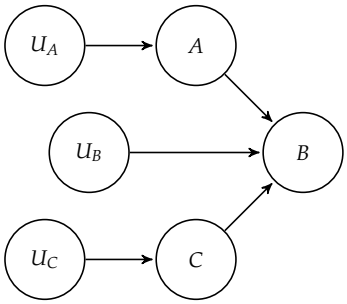
the plausibility relation. Let us start with the second example from that section, repeated below.

**3.** EXAMPLE (**Revisiting Three Sisters with causality**). *Consider the case of three sisters who own just one bed, large enough for two of them but too small for all three. Every night at least one of them has to sleep on the floor. However, whenever Ann and Carol are both in bed, they will kick off Billie; she will have to sleep on the floor.*

*Suppose Carol is invisible. Suppose further that you are a proud parent of Ann, Billie and Carol. Before you go to bed you go in and check on the kids. As described in the original version, Ann is on the floor, Billie is in bed and Carol (obviously) is also in bed. Now you turn to your spouse and comment:*

(4) If Ann had been in bed, Carol would have been on the floor.

Let $A = 1, B = 1$ and $C = 1$ stand for "Ann is in bed", "Billie is in bed" and "Carol is in bed" and let $U_A$, $U_B$ and $U_C$ be the external factors that decide whether Ann, Billie and Carol want to sleep on the floor or in bed. Since in the story Ann and Billie are visible while Carol is invisible, the information partition is defined by the values of the variables $A$ and $B$. Figure 3.7 represents the relevant causal structure for this example. This causal structure will restrict the plausibility order relevant for this example. But there is a additional, non-causal law that also restricts the plausibility order in the given context: because there are only two available beds, in the most plausible worlds there will be two girls in bed and the third one on the floor. Thus, other things being equal, worlds in which exactly two girls are in bed will be preferred to worlds where less or more girls are in bed. Because there are no other restrictions mentioned in the context, we assume that the plausibility order is the weakest order satisfying these restrictions.




| $F_A$: | $U_A$ | $A$ |
|---|---|---|
| | 0 | 0 |
| | 1 | 1 |

| $F_C$: | $U_C$ | $C$ |
|---|---|---|
| | 0 | 0 |
| | 1 | 1 |

| $F_B$ : | $U_B$ | $A$ | $C$ | B |
|---|---|---|---|---|
| | 0 | 0 | 0 | 0 |
| | 0 | 0 | 1 | 0 |
| | 0 | 1 | 0 | 0 |
| | 0 | 1 | 1 | 0 |
| | 1 | 0 | 0 | 1 |
| | 1 | 0 | 1 | 1 |
| | 1 | 1 | 0 | 1 |
| | 1 | 1 | 1 | 0 |

Figure 3.7: The causal dependencies in the three sisters example.

Now, let us check the truth conditions of the counterfactual $Bel[A = 1]Bel(C = 0)$ given this model and the observation that Ann is on the floor and Billie is in bed. The most plausible world given the available information is a world $w$ in which Ann doesn't sleep in bed, Billie does, and (because of the preference for two girls in bed) Carol does as well (see $w$ in Figure 3.8). Thus, we have to check whether $M, w \models [A = 1]Bel(C = 0)$, i.e. $M, w_{A=1} \models Bel(C = 0)$, where $w_{A=1}$ is the world whose setting of exogenous variables is same as $w_1$, $A$ is set to 1 and all other endogenous variables are recalculated from this and the causal laws. In particular, because $B$ causally depends on $A$ (and $C$) the value of this variable is recalculated and now is set to 0 (see $w_{A=1}$ in Figure 3.8). To check whether $Bel(C = 0)$ holds in this world we have to check whether $C = 0$ holds in the preferred worlds in this cell of the partition. This the world $w'$ in Figure 3.8, where Ann and Carol are in bed, but Billie isn't. Thus, $C = 0$ doesn't hold in this case. Thus, the counterfactual is predicted to be false, as intended.

Two interesting aspects of our approach can be observed in this context. First of all, notice the dominance of causal reasoning over epistemic reasoning in our approach. First, the causal impact of the intervention is calculated, resulting in concluding that Billie can no longer be in bed. Only then are epistemic consequences considered. This is also one of the many examples where an analysis that does not assume an extra epistemic operator in the consequent would have made the same prediction. Both analyses, (5b) and (5c), agree that (4) needs to be rejected in this context.

**Three sisters – version 1.** However, in Section 3.6.1 we also discussed a different version of the context in which the counterfactual *If Ann had been in bed, Carol would have been on the floor* was acceptable. This version, Example 4, didn't contain any information about causal links between the sleeping places of the three sisters. In other words, the causal model now looks as in Figure 3.7,

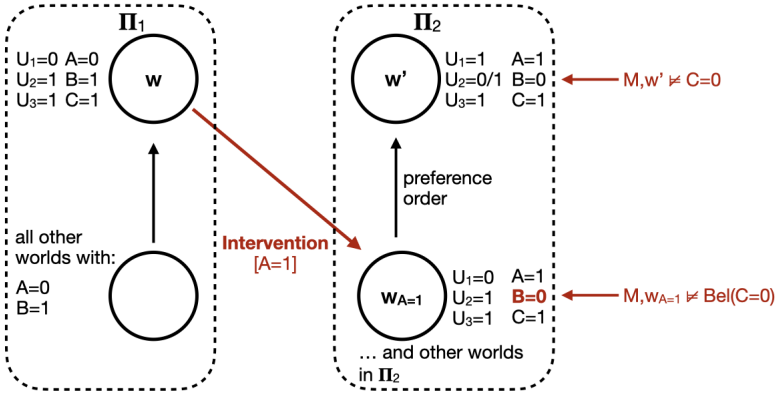



Figure 3.8: A graphic representation of the causal epistemic model of the three sisters Example.

but without any arrows connecting the variables $A$, $B$ and $C$ (consequently, $F_B(B)$ only depends on $U_2$). To check the predictions our approach makes for (4) in this context, we have to check whether $[A = 1]Bel(C = 0)$ holds with respect to this model and the the world $w$ where Ann is on the floor and Billie and Carol are sleeping in the bed (see $w$ in Figure 3.8). But now $w_{A=1}$ will still have Billie sleeping in bed, because changing the value of Ann will not causally affect the whereabouts of Billie. We end up in a different cell of the partition (one where Ann and Billie are observed to be in bed), and the most preferred worlds in this cell have Carol sleeping on the floor. In the world we reach after intervention, we would *see* Ann and Billie sleep in the bed and conclude that Carol must be on the floor. The counterfactual comes out as acceptable, just as intended.

## 3.6 Discussion

### 3.6.1 Comparing with an approach using belief revision

In Section 3.3.2 we argued against an approach that distinguishes next to a causal reading of counterfactuals also an epistemic reading based on belief revision. Some readers might have observed that the examples discussed in the previous section (the King Ludwig example and the two versions of the example with the three sisters) could also been explained using this approach, proposing a reading based on the particular version of belief revision introduced in Section 3.3.2. In Section 3.3.2 we argued against this approach based on problems with disambiguating between the two readings. Now we can add also empirical reasons to prefer our approach. A first type of evidence comes from counterfactuals that involve a combination of causal and epistemic

reasoning. Consider, for instance, the following extension of the three-sisters example.

**6.** Example. **(Four Sisters)**
*Consider the case of four sisters (Ann, Billie, Carol, and Dotty) who own two beds, one red and one blue. Both beds are large enough for two of the girls but too small for three or four. As one can imagine, four girls living together in one room leads to some tension and Ann really hates Billie and Dotty at the moment. Therefore, whichever bed Ann chooses, she will kick off Billie and Dotty; they have to sleep in the other bed.*

*Suppose Carol is invisible. Suppose further that you are a proud parent of Ann, Billie, Carol and Dotty. Before you go to bed you go in and check on the kids. You see that Ann is in the red bed, and Billie and Dotty are in the blue bed. Carol (obviously) is also in the red bed. Now you turn to your spouse and comment:*

(6) If Ann had been in the blue bed, Carol would have been in the blue bed as well.

This sentence is intuitively true in the given context. However, to predict this judgement, one needs to combine causal and epistemic reasoning. In this new example Ann's sleeping choices will causally determine where Billie and Dotty sleep. At the moment Ann is sleeping in the red bed. If you change this by intervention to the blue bed, then this will cause Billie and Dotty to move to the red bed (in the world resulting from intervention, Carol would be in the red bed as well). But given what you would observe in this counterfactual scenario (Ann in the blue bed, Billie and Dotty in the red bed), you would conclude that Carol is with Ann in the blue bed (this would be the most preferred world. An ambiguity approach along the lines sketched in Section 3.3.2 would not be able to deal with examples like this, because according to such an approach there is always a decision that needs to be made between either causal reasoning or epistemic reasoning. Both modes of reasoning cannot be combined. Notice additionally that we need the particular order in which both modes of reasoning are applied in our framework: first causal reasoning based on intervention, then epistemic reasoning on the result.

The following epistemic variation of the circuit example illustrates the differences between the kind of epistemic inferences our approach gives rise to and what belief revision would predict.

**7.** Example. **(Circuit with invisible switch)**
*Suppose there is a circuit such that the light is on (L) exactly when both switches are in the same position (up or not up). The switches are at both ends of a long corridor. You are standing at one end close to switch 1. You can see this switch and the lamp, but you can't see the other switch. At the moment switch 1 is down ($\neg S_1$), the lamp is out ($\neg L$). Thus, switch 2 is (obviously) up ($S_2$).*

(7) If switch 1 had been up, the light would have been on.

Intuitively, this counterfactual is true in the given context. The proposal involving belief revision that we discussed in Section 3.3.2, however, predicts the counterfactual (7) to be false. Because the state of the lamp can be observed, this state is fixed in the process of belief revision and the position of the second switch (which can't be observed) is varied. Thus, the approach predicts that if switch 1 had been up, the light would still have been on, but the second switch would have been down. Our approach still predicts the counterfactual to be true. Again, the reason is the dominance of causal reasoning over epistemic reasoning. A defender of the belief revision account could now argue that in this case the counterfactual should get a causal reading instead of an epistemic one. But then we are back at the point we already made in Section 3.3.2: why should the sentence get a causal reading and not an epistemic one?

### 3.6.2 Backtracking

Another interesting test case for our approach is backtracking. Backtracking conditionals, as we understand them here, reason from effect to cause.[21] The general position in the philosophical literature is that while backtracking is completely fine for indicative conditionals, this is not the case for counterfactuals. Lewis [1979] claims that backtracking counterfactuals can be acceptable, but only with a lot of pushing from the context. A similar observation can be found in Frank [1997] and many other papers. Backtracking also became an important issue in the discussion of causal approaches to counterfactuals. A causal approach based on intervention, like the one we discussed in Section 3.2 excludes backtracking counterfactuals. This prediction has been tested in a number of experiments with different results. While the majority of studies seem to confirm the predictions of the causal approach [Sloman and Lagnado, 2005, Gerstenberg et al., 2013], there where also some contradicting findings [Rips, 2010, Rips and Edwards, 2013, Dehghani et al., 2012].

Our approach makes some rather specific predictions concerning backtracking. If all variables are observable, backtracking is not possible, just as in standard causal approach to counterfactuals. If, however, the counterfactual reasons from effect to cause and the cause is not observed, backtracking can occur. This is, in fact, exactly what happens in the King Ludwig example. But we can also illustrate this point with the following example.

[21] It is also possible to define backtracking as a temporal property of counterfactuals: the eventuality described in the antecedent takes place after the eventuality described in the consequent. We chose the causal definition, because causal dependencies are the focus of this paper and it allows us to circumvent debates concerning the relation between temporal order and causation that we would have to dive into otherwise.

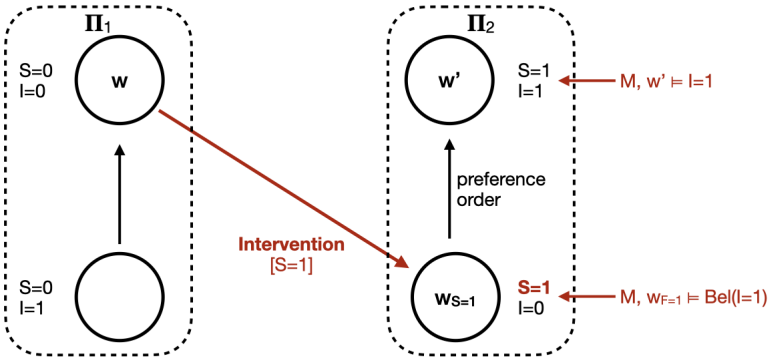



Figure 3.9: Interpreting (8)

**8.** EXAMPLE. **(Interview)**
*Susy and Mary both have a very important job interview today. Mary goes in first, Suzy is waiting outside the interview room. When Mary comes out she looks rather unhappy. Suzy thinks:*

*(8) If Mary had left the interview smiling, the interview would have gone well.*

(8) is one of the few backtracking examples in the literature with unobservable variables in the consequent. Our approach predicts the backtracker to be acceptable in this context. The reasoning behind this prediction is basically the same as in the King Ludwig example. Let $S$ stand for *Mary comes out smiling* and $I$ for *the interview went well*. We assume a causal relation between $I$ and $S$ such that $I$ causes $S$ to hold. Given the observation that Mary comes out unhappy, the speaker believes that she is in a world where $I$ and $S$ both don't hold (see world $w$ in Figure 3.9), because this is the most plausible world given her observations. Intervention with $[S = 1]$ will bring us to a world $w_{S=1}$ where the interview still went badly, but Mary came out smiling. But in this case the agent would have believed that the interview went well, i.e. that the actual world is $w'$ in Figure 3.9. Thus the counterfactual is predicted to be true.

Unfortunately, so far no experimental studies on backtracking have been conducted that involve unobservable variables. So, there is no data that we can test our approach against. This is something we hope to amend in future work. These studies might report that people find the backtracker (8) unacceptable – contradicting our approach. There are two possible explanations for such an observation that are still consistent with our approach. First of all, we could argue that $I$ is a variable that should be counted among the observable variables in this case. Taking the knowledge partition to be always based just on direct observation is not plausible. There might be other sources of evidentiality that speakers also take into account when differentiating between information that

they take for granted and information that is less solidly grounded. One might argue that this example uses a different categorisation for what counts for the information partition and what counts for the plausibility order. But that won't be easy. Whatever story one might come up with here, this story also has to explain why in the interview example the unobservable *I* should count for the definition of the information partition, while the equally unobservable *K* in the King Ludwig example should not.

A second solution could be to go back to an ambiguity approach to counterfactuals and propose a second reading for counterfactuals. Concretely, the second reading we would need to propose would be the original causal reading, i.e. reading (5b) on page 54. But, as discussed above, if one chooses this option, one needs to also have a story explaining how disambiguation between the two readings works. The situation might be less dire in our case than for the ambiguity approach discussed in Section 3.2. In our case the difference between the two readings is smaller. We could say that disambiguation works based on which question about the introduced hypothetical scenario we want to see answered. But still some story about disambiguation needs to be told. However, before we need to consider these options to save our approach, there first has to be serious evidence that the approach as it is now is on the wrong track.

### 3.6.3 Epistemic counterfactuals without causal information

To conclude this paper let us discuss an example that seems to pose a real challenge for the approach we defend here: the Treasure Hunt example from Edgington [2011].[22]

**9.** Example. **(Treasure Hunt)**
*Ali and Bob are playing as a team in a treasure hunt, with on prize. A parent gives them a hint: The prize is in the attic or the garden. Ali says to Bob: "You go check the garden. I'll search the attic." 10 minutes later Ali shouts: "I found the price in the attic!" Bob asks Ali: "Well, then why did you tell me to search in the garden?" Ali replies: "Because ...*

*(9) ... if the prize hadn't been in the attic, it would have been in the garden."*

This example bears the same structure as many other famous examples of counterfactuals in the literature. We first get disjunctive information (the treasure is either in the garden or in the attic). Later we find out which of the disjuncts holds. Now, we are asked to consider counterfactual reasoning of the form *if it hadn't been this disjunct that was true, it would have been the other*. In some of the cases discussed in the literature, like the famous example involving the

[22] Edgington borrowed and adapted the example from Grice [1975].

assassination of Kennedy, there is strong consensus that the counterfactual is not acceptable.[23] In other examples, like the example involving a man eating a Hamburger from Hansson [1989] the judgements have been less clear. But in case of the Treasure Hunt the consensus in the literature is that the relevant counterfactual (9) is acceptable in the given context (though this should be confirmed in experimental work).

Accounting for such an observation is challenging for our approach. In general, in a situation with three variables $A$, $B$ and $C$, where $A$ and $B$ are alternative, independent causes of $C$ and we observe $A$ and $C$ to be the case, while $B$ is observed not the hold, we predict the counterfactual *If A had not been the case, then B would have been the case* to be false. And that is also exactly the prediction we want for the Kennedy example and many similar examples. But in case of the Treasure Hunt example counterfactual reasoning from the falsity of one disjunct to the truth of the other seems to be possible. The challenge is not only to account for the acceptability of (9) in this case, but at the same time to also explain why the same type of reasoning is not acceptable in these other examples.[24] There needs to be differences between the examples that the proposed semantics for counterfactuals is sensible to. Indeed, the epistemic structure we added to the causal approach to counterfactuals allows us to model the Treasure Hunt example, but still explain what is special about this example in contrast to all the other cases where this type of counterfactual reasoning is not acceptable.

Let us have a look at how we can model the given context. There are at least two endogenous variables in the relevant epistemic causal model: $A$ for the treasure is in the attic, $G$ for the treasure is in the garden. These variables are not causally dependent on each other. Given that Ali did find the treasure in the attic, $A$ is a variable that has been observed to be true. Bob did not find the treasure in the garden. But he also didn't observe the treasure to not be in the garden, that is something he inferred from Ali finding the treasure in the attic. Thus, $G$ is not observable. This is a first important assumption we make about the example: not everything is observed. This makes room for epistemic reasoning to the place.

The second important assumption concerns the preference order and the way we interpret the disjunctive information about the treasure being either in the garden or in the attic. In this particular scenario, this is not just factual information, but actually defines the game that Ali and Bob are playing with

[23] The duchess example from Veltman [2005] is a more neutral version of this example. Also here the counterfactual is judged to be not acceptable.

[24] An approach in terms of belief revision would struggle with the second part. There is an easy way to explain (9) in terms of belief revision, but what then about the Kennedy case or the duchess example? We could, of course, fall back again on claiming that these counterfactuals get a causal reading. But that brings us back to the point we made before: what decides which counterfactual gets which reading?

their parents, i.e. it informs the plausibility relation: worlds where the treasure is in either of the two locations are preferred above all other worlds. This results in the model given in Figure 3.10. Now, we can check the acceptability of the counterfactual in (9). Intervention with its antecedent of will bring us to a world $w_{A=0}$ in which the treasure is neither in the attic nor in the garden. But given that $A$ is observable, the agent would have believed in this case that the treasure is in the garden. Thus, the counterfactual is predicted to be acceptable in the given context.

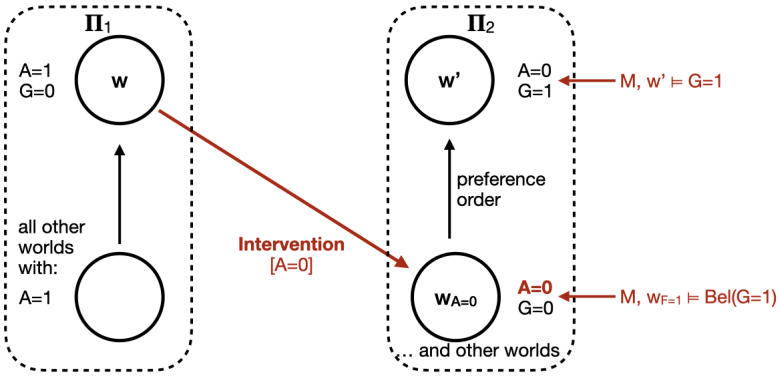



Figure 3.10: The interpretation of (9) in the context of the Treasure Hunt example.

We think that this example illustrates nicely the value added by the epistemic component in our approach. What does the work here is the plausibility relation (which does not only consider causal laws, but also all kinds of other belief preferences we might have) and the distinction of observable and unobservable variables. Adding these epistemic components allows us now to see differences between examples that so far have all been sorted into the same box. It also allows us to account for differences between judgements of speakers. What you take to be solid knowledge or defeasible belief might vary between different agents in the same context. The same holds for what agents take to inform their plausibility order. This might explain the variation in judgements we often observe.

## 3.7 Conclusions

This paper focused on counterfactuals that seem to involve epistemic reasoning. We saw that that a standard causal approach to counterfactuals based on intervention is not able to account for these examples. We also discussed an approach distinguishing a second, epistemic reading for counterfactuals based on belief revision. This approach struggles with explaining the disambiguation between the two readings, but in later sections we also discussed empirical

problems for this line of approach. Following these insights we then proposed an account for counterfactuals that extends a causal, interventionist approach with epistemic reasoning. Counterfactuals are described as reasoning about what an agent would believe given some intervention. This approach allows us to explain the acceptability of Kratzer's King Ludwig example and also other related epistemic counterfactuals. The main innovation of our approach is that it allows us to account for causal and epistemic examples of counterfactual with one single interpretation rule for counterfactuals; we do not need to distinguish different readings.

We discussed how the approach deals with a number of key examples. But there are many more examples in the literature that we did leave out to keep the length of the paper manageable. There are also many possible extensions and interpretations of the approach we still need to look into in future work. One issue is, for instance, the way we distinguished here between knowledge and belief. We based knowledge on observables and added belief based on laws. But there could be other sources of information that we might want to consider as contributing to knowledge or belief. Furthermore, it is worth considering whether we really want to have this clear distinction between either knowledge or belief, or whether we want to model the distinction more gradually. Given the approach we proposed here, one might also wonder how the the meaning of counterfactuals relates to the more general debate about the distinction between knowledge and belief in the philosophical literature or the linguistic debate concerning evidentiality. Another direction for future research is the question how the approach proposed here fits with linguistic properties of counterfactuals. For instance, can it be combined with a compositional approach to their meaning? And how does the analysis proposed here relate to an approach to indicative conditionals?

To formulate our approach to the meaning of counterfactuals, we introduced an epistemic extension of causal models. Also with respect to this type of combination of causal reasoning with epistemic reasoning there are still many open questions. One possible direction of future research could be to think about how to model causal dependence between facts and beliefs. An application where this might become relevant is the semantic analysis of causal verbs like *made* or *cause*. Nadathur and Lauer [to appear] recently made an interesting proposal for the semantics of these verbs using causal models. But this approach works entirely within a causal framework, without any epistemic component. This will be not sufficient if one also wants to account for statements like *you made me believe that ...* or *that caused me to think that ....* We see two possible ways to address this problem. We could use our indirect model of the interaction between causal and epistemic reasoning and provide a semantics for these verbs that goes beyond direct causal dependence, contra to what Nadathur and Lauer [to appear] propose. Or we could enrich the model with a representation of causal dependencies between certain variables

and, for instance, the information partition. Then we could keep a completely causal analysis of the semantics of these verbs. We leave it to future work to decide between these two options.

# Chapter 4

# A Logical and Empirical Study of Right-Nested Counterfactuals

## 4.1 Introduction

The meaning of counterfactual conditionals bears an intrinsic relation to a number of central scientific problems, like the nature of reasoning, the possibility of knowledge, and the status of laws of nature. Therefore, this topic has fascinated many thinkers from various disciplines, like philosophy, logic, psychology and others. But, despite a lot of effort, no consensus has been reached yet about how the meaning of these sentences needs to be approached.

One way to conceptualize the evaluation of counterfactuals, very common in the literature, goes as follows. When evaluating a counterfactual, we select, given the antecedent $A$ and the context of evaluation, certain (hypothetical) situations in which the antecedent is true, and then check whether they make the consequent $B$ true as well. The challenge of accounting for counterfactuals then becomes to define the relevant *selection function* correctly. Following the approach of Stalnaker and Lewis Lewis [1973], Stalnaker [1968], which still is the dominant approach in the philosophical literature, the selection is based on similarity: we take those hypothetical situations that are most similar to the actual world. But this proposal is known to be problematic: among other things, it appears to be too flexible. In recent years, the interventionist approach to counterfactuals became very popular (Pearl [2013], Schulz [2011], Kaufmann [2013], Halpern [2016], Ciardelli et al. [2018] and others). This approach describes the truth conditions of counterfactuals with respect to a representation of the relevant causal dependencies, building on Causal Models as introduced in Pearl [2000], Spirtes et al. [2000]. The approach got its name from the way it describes the selection function. In the selected hypothetical scenarios, the antecedent has been made true by intervention on the actual causal dependencies: it is cut off from its causal parents and stipulated to be true by law. Then,

one checks in the resulting model whether the consequent holds.[1]

Recently, this approach has been criticized by Fisher Fisher [2017]. He claims that interventionism makes incorrect predictions for right-nested counterfactuals. According to Fisher, the problem is a particular property of the interventionist approach, *strict interventionism*, which he argues needs to be dropped in a proper account. We will argue, using the results of an empirical study, that Fisher is right in his critique. But this does not mean that the interventionist approach needs to be given up in general. We will propose a variation of the approach that drops strict interventionism and thus can account for Fisher's core-observations. We will also make precise how this new proposal relates to the classical interventionist approach as spelled out in Halpern [2016]. We will do so by providing an axiomatization of the new operator for counterfactual reasoning that we introduce. As it will turn out, this new operator can be already defined in terms of the classical intervention operator. Furthermore: to a large extend, they both make the same counterfactuals true. So our proposal, though formalizing a slightly different take on what intervention means, is in terms of logical properties a conservative change of the original interventionist approach.

## 4.2 The interventionist approach to counterfactuals

Our presentation of the interventionist approach to counterfactuals is based on the one proposed by Briggs in Briggs [2012]. Still, we will only introduce the parts that are relevant for the discussion at hand. The two central ingredients of the approach are **(i)** the causal model, which contains information about the relevant causal dependencies, and **(ii)** the operation of intervention, which defines the selection function by mapping a given causal model onto a class of models that make the antecedent of the relevant counterfactual true.

Causal models represent the causal dependencies between a given finite set of variables. For each variable $X$ we fix its range $\mathcal{R}(X)$, the set of possible values the variable can take. The variables are sorted into the set $\mathcal{U}$ of *exogenous* variables (those whose value is independent from the value of other variables in the system) and the set $\mathcal{V}$ of *endogenous* variables (those whose value causally depends on the value of other variables in the system). Based on this distinction, a causal model can be defined as given in Definition 18.

**18.** Definition (Causal model). *Let $\mathcal{U} = \{U_1, \ldots, U_m\}$ be the set of exogenous variables, and $\mathcal{V} = \{V_1, ..., V_n\}$ be the set of exogenous variables. A* causal model *over $\mathcal{U}$ and $\mathcal{V}$ is a tuple $\langle \mathcal{S}, \mathcal{A} \rangle$, defined as follow.*

[1] It turns out that for recursive causal models the interventionist selection function can be understood as just one particular way to make similarity precise Halpern [2013], Marti and Pinosio [2014b].

- *The first component, $\mathcal{S}$, is a set $\{f_{V_j} \mid V_j \in \mathcal{V}\}$ assigning to each endogenous variable $V_j$ a map*

$$\Big(\mathcal{R}(U_1)\times\cdots\times\mathcal{R}(U_m)\times\mathcal{R}(V_1)\times\cdots\times\mathcal{R}(V_{j-1})\times\mathcal{R}(V_{j+1})\times\cdots\times\mathcal{R}(V_n)\Big) \to \mathcal{R}(V_j).$$

- *The second component, $\mathcal{A}$, is the* valuation *function, assigning to every $X \in (\mathcal{U} \cup \mathcal{V})$ a value $\mathcal{A}(X) \in \mathcal{R}(X)$ that* complies with the structural functions in $\mathcal{S}$*: for all endogenous variables $V_j \in \mathcal{V}$, we have*

$$\mathcal{A}(V_j) = f_{V_j}\Big(\mathcal{A}(U_1),\ldots,\mathcal{A}(U_m),\mathcal{A}(V_1),\ldots,\mathcal{A}(V_{j-1}),\mathcal{A}(V_{j+1}),\ldots,\mathcal{A}(V_n)\Big).$$

In a causal model, the set $\mathcal{S}$ fixes the causal dependencies among the variables. Each map $F_V$, sometimes called $V$'s *structural function*, describes how the value of $V$ causally depends on that of other variables. The function $\mathcal{A}$ defines the value of all variables in the model. It does so in a way that is consistent with the causal laws fixed by $\mathcal{S}$. Here are two causality-related concepts that will be important in the rest of the text.

**19.** Definition (Dependency). *Let $\langle \mathcal{S}, \mathcal{A}\rangle$ be a causal model over $\mathcal{U} = \{U_1, ..., U_m\}$ and $\mathcal{V} = \{V_1, ..., V_n\}$. Given an endogenous variable $V_j \in \mathcal{V}$, let $\langle X_1, \ldots, X_{m+n-1}\rangle$ be the (m+n-1)-tuple $\langle U_1, \ldots, U_m, V_1, \ldots, V_{j-1}, V_{j+1}, \ldots, V_n\rangle$.*

*We say that the endogenous variable $V_j \in \mathcal{V}$ is* directly dependent *on a variable $X_i \in (\mathcal{U} \cup \mathcal{V}) \setminus \{V_j\}$ (in symbols, $X_i \hookrightarrow_{\mathcal{S}} V_j$) if and only if there is $x_1 \in \mathcal{R}(X_1)$, $\ldots, x_{i-1} \in \mathcal{R}(X_{i-1})$, $x_{i+1} \in \mathcal{R}(X_{i+1})$, $\ldots, x_{m+n-1} \in \mathcal{R}(X_{m+n-1})$) and there are $x_i' \neq x_i'' \in \mathcal{R}(X_i)$ such that $F_{V_j}(x_1, \ldots, x_i', \ldots, x_{m+n-1}) \neq F_{V_j}(x_1, \ldots, x_i'', \ldots, x_{m+n-1})$. When $X_i \hookrightarrow_{\mathcal{S}} V_j$, we will also say that $X_i$ is a* parent *of $V_j$.*

*We say that $V_j \in \mathcal{V}$ is* causally dependent *on $X_i \in (\mathcal{U} \cup \mathcal{V}) \setminus \{V_j\}$ if and only if $X_i \hookrightarrow^+_{\mathcal{S}} V_j$, with $\hookrightarrow^+_{\mathcal{S}}$ the transitive closure of $\hookrightarrow_{\mathcal{S}}$.*

**20.** Definition (Recursive causal models). *A causal model is said to be* recursive *if and only if $\hookrightarrow^+_{\mathcal{S}}$ is a strict partial order (so there are no circular dependencies between the variables).*

A recursive model is a model in which no circular causal dependencies occur. In a recursive causal model $\langle \mathcal{S}, \mathcal{A}\rangle$, if the values of all exogenous variables are fixed, the value of every endogenous variable $V$ is uniquely determined (from the values of the exogenous variables and the causal dependencies as described by $\mathcal{S}$). In the rest of the paper we will only consider recursive causal models, which from now on will be called simply *causal models*.

We use a simple propositional language extended with an operator for counterfactual conditionals to talk about causal models.

**21.** Definition (Language $\mathcal{L}_{[]}$). *The set of formulas $\phi$ of the language $\mathcal{L}_{[]}$ over $\mathcal{U} \cup \mathcal{V}$ is defined by*

$$\phi ::= X{=}x \mid \neg\phi \mid \phi \wedge \phi \mid [\overrightarrow{X}{=}\overrightarrow{x}]\phi \quad \text{for } X \in \mathcal{U} \cup \mathcal{V},\ x \in \mathcal{R}(X),\ \overrightarrow{X} = (X_1, \ldots, X_k) \in (\mathcal{U} \cup \mathcal{V})^k,\ k \in \mathbb{N},\ X_i \neq X_j \text{ for } i \neq j,\ \overrightarrow{x} = (x_1, \ldots, x_k) \text{ with } x_i \in \mathcal{R}(X_i)$$

Thus, $\mathcal{L}_{[]}$ extends the basic causal language of Halpern [2016] by allowing right-nested counterfactuals. Still, it is only a fragment of the language used in Briggs [2012], as it does not allow Boolean combinations of atoms in the antecedent (which are not relevant to the discussion here). Sentences of the form $[\overrightarrow{X}{=}\overrightarrow{x}]\phi$ should be read as *"if the variables in $\overrightarrow{X}$ were to be set to $\overrightarrow{x}$, then $\phi$ would hold"*, with the variables in $\overrightarrow{X}$ called the *intervened* variables. These kind of formulas are taken to represent counterfactuals. In contrast to most literature on causal models, our definition allows for exogenous variables in the antecedent of a counterfactual.

The second important ingredient of the interventionist approach is the notion of intervention involved in the interpretation rule for formulas of the form $[\overrightarrow{X}{=}\overrightarrow{x}]\phi$. This is the way the selection function is defined in the case of this particular approach to counterfactual conditionals. Given a causal model $\langle \mathcal{S}, \mathcal{A} \rangle$ and a counterfactual sentence $[\overrightarrow{X}{=}\overrightarrow{x}]\phi$, intervention provides a set of models satisfying the antecedent $\overrightarrow{X}{=}\overrightarrow{x}$. For these models we will then check whether the consequent $\phi$ holds as well. The models are constructed in two steps. First, the variables in $\overrightarrow{X}$ are forced to the values assigned by the antecedent $\overrightarrow{X}{=}\overrightarrow{x}$. In the case of exogenous variables, this is done by simply changing their value indicated by $\mathcal{A}$; in the case of endogenous ones, this is done by turning the variable's structural function in $\mathcal{S}$ into a constant function. These variables become effectively exogenous variables.[2] Then, the values of the endogenous variables are calculated using the new structural functions.

**22.** Definition (Intervention). *Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a causal model. When evaluating formulas in $\mathcal{L}_{[]}$, the semantic interpretation of Boolean operators is as usual; for the rest,*

$$\begin{array}{lll} \langle \mathcal{S}, \mathcal{A} \rangle \models X = x & \textit{iff}_{def} & \mathcal{A}(X) = x \\ \langle \mathcal{S}, \mathcal{A} \rangle \models [\overrightarrow{X}{=}\overrightarrow{x}]\phi & \textit{iff}_{def} & \langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}} \rangle \models \phi \end{array}$$

*with $\langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}} \rangle$ the causal model where*

[2] In a setting that allows for disjunction in the antecedent of counterfactuals (e.g., Briggs [2012]), the previous steps might produce more than one causal model. However, for the possible antecedents considered in our language, the resulting model is uniquely defined.

(i) *$\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}$ is as $\mathcal{S}$ except that, for each endogenous variable $X_i$ in $\overrightarrow{X}$, the function $f_{X_i}$ is replaced by a 'constant' function $f'_{X_i}$ that assigns to $X_i$ the value $x_i$ regardless of the values of all other variables.*

(ii) *$\mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}$ is the unique assignment to causal variables that is identical to $\mathcal{A}$ with respect to exogenous variables not in $\overrightarrow{X}$, assigns to each exogenous variable $X_i$ in $\overrightarrow{X}$ the indicated value $x_i$, and complies with the causal dependencies in $\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}$ for the endogenous ones.*[3]

Thus, according to the interventionist approach, the selection function $f$ discussed in the introduction should be defined as

$$f\left(\langle \mathcal{S}, \mathcal{A}\rangle, \overrightarrow{X}=\overrightarrow{x}\right) := \langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}\rangle.$$

It is worthwhile to emphasise that, in the intervened model $\langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}\rangle$, the valuation $\mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}$ *complies* with the model's causal dependencies, $\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}$: for every $V \in \mathcal{V}$, the value $\mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}(V)$ is given by the variable's structural function as provided by $\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}$. So, intervention happens at the level of causal dependencies, and this change affects the valuation $\mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}$. In Section 4.5 we will introduce a notion of intervention that, working on a more general class of models, changes values directly, leaving causal dependencies unaffected.

## 4.3 Fisher's criticism

Fisher (Fisher [2017]) criticizes the approach described above. He claims that it makes incorrect predictions for right-nested counterfactuals. Concretely, he discusses the examples (10) and (11) below.[4]

- **Match.** I hold up a match and strike it, but it does not light. I say

  (10) If the match had lit, then (even) if it had not been struck, it would have lit.

[3] Note: the assignment is unique, not only because the values of exogenous variables is determined, but also because, if $\langle \mathcal{S}, \mathcal{A}\rangle$ is recursive, so is $\langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}\rangle$. This is because the intervention operation only removes causal dependencies, and thus no circular dependencies are added.

[4] Fisher also considers another example, involving the counterfactual "If the match were struck and it lit, then if it hadn't been struck, it would have lit". This is not a good example to make his point, as it contains a conjunction of cause (striking the match) and effect (the match lights) in the antecedent. For the counterexample to work, Fisher needs this conjunction to be interpreted as two independent interventions. However, it could be that *"and"* is interpreted causally in this case: *"If the match were struck and because of that it lit, ..."*. But then the fact that the match lights would be introduced as a causal consequent of the striking of the match and not as an independent intervention.

- **Headlamp.** I hold up a headlamp in good working condition. I say

  (11) If the headlamp were emitting light, then if it had had no batteries, the headlamp would be emitting light.

Both examples involve a model of the form shown in Figure 4.1 left side, where $A_1$ stands for the variable the first antecedent talks about and $A_2$ for the variable of the second antecedent.[5]


Figure 4.1: A causal model for **Match** and **Headlamp**, before (let side) and after (right side) interpreting the counterfactuals.

Following the interventionist approach, the evaluation of the first antecedent produces a causal model where $A_1$ is forced to a particular value, and where the causal connection between $A_2$ and $A_1$ has been erased. Evaluating the second antecedent forces $A_2$ to a particular value too, but this will no longer affect $A_1$. Hence, the counterfactuals (10) and (11) are predicted to be true, but intuitively, according to Fisher, they should be false. Fisher traces the problem back to the property of *strict interventionism* (SI).

> (SI) "When a variable $V$ is intervened on so that it is made to take a value $v$, $V$ remains set to $v$ unless it is intervened upon again per an iterated application of the interventionist recipe." (Fisher [2017]:4939).

Interventionist approaches have this property because their selection function maps a given causal model $M$ and an antecedent $A$ to a new causal model in which a causal variable $V$ occurring in the antecedent $A$ has lost all connections to its causal parents. Any later intervention that might affect $V$'s (former) causal parents will no longer affect $V$ itself. So, as long as $\psi$ does not assign a new value to $V$, the counterfactual $[\overrightarrow{V}=\overrightarrow{v}][\psi](V_i = v_i)$ will always come out as true.

To solve this problem, Fisher proposes that we have to give up strict interventionism. More concretely, he proposes the following adequacy condition for approaches to the meaning of counterfactuals: "A causal model semantics for counterfactuals should admit cases in which the variables implicated in the antecedent of a counterfactual remain causally sensitive to their parents throughout the evaluation procedure." (Fisher [2017]:4942). However, he does

[5] We ignore other possible variables, as they will not affect the relevant predictions made.

not propose an alternative approach that has this property.[6] In the rest of the paper we will do the following. First of all, we need to confirm Fishers judgements concerning the target examples (10) and (11) with an actual survey. These are not your every day examples of counterfactuals and we should make sure that the intuitions Fisher reports are generally shared and that they really concern truth conditions, not the assertability of this type of right-nested counterfactuals. This is the subject of Section 4.4. In Section 4.5 we will develop an alternative interventionist approach to the meaning of counterfactual conditionals that is not strictly interventionist. Finally, in Section 4.6 we will extend the discussion with some additional examples and investigate whether giving up strict interventionalism is sufficient to account for right-nested counterfactuals in general.

## 4.4 An empirical study on Fisher's counterexamples

A possible objection against Fisher's observations and the conclusions he derives from them is that he confuses judging a sentence false with rejecting it as not well-formed. Maybe we are inclined to say "No" to the counterfactuals in (10) and (11), because they are very strange counterfactual sentences. To exclude this interpretation of the observations, we conducted a small empirical study in which we asked the participants to judge not only the counterfactuals (10) and (11), but also their counterparts (12a) and (12b) in which the final consequent has been negated. If participants judge the sentences (10) and (11) false because they consider them defective, they should judge (12a) and (12b) to be defective (and hence false) as well.

(12) a. If the match had lit, then if it had not been struck, it would not have lit.

b. If the headlamp were emitting light, then if it had had no batteries, the headlamp would not have been emitting light.

### 4.4.1 Method & Participants

We used the scenarios **Match** and **Headlamp** from page 73 and a third scenario containing a counterfactual $[\varphi][\psi]\xi$ with $\xi$ talking about a causal effect of $\varphi$. For each scenario we asked the participants to judge 3 counterfactuals: the target right-nested counterfactual, the counterfactual with the opposite final consequent and a filler item to check whether the participants where paying

[6] Fisher discusses in Fisher [2017] an alternative definition of intervention, dubbed "side-constrained intervention", but admits that this variation is not really targeting the root of the problem.

attention and understood the presented scenario correctly. This resulted in 9 questions that the participants had to answer. The order of question was randomized. The participants had to judge the truth value of the counterfactual using a slider bar with five values, from 0 to 4. They were told that 0 means the sentence is false, 4 it is true and 2 that the truth value is unclear. The values 1 and 3 allowed them to indicate that they find a sentence weakly false or true.

The study was implemented in Qualtrics, a web-based survey tool. Participants were recruited via Prolific.ac, an online platform aimed at connecting researchers and participants willing to fill in surveys and questionnaires in exchange for compensation for their time (Palan and Schitter [2018]). We recruited native English speakers (British and American English). Fifty-two participants completed the task. Eight participants were excluded. Two participants did not answer the filler question for the match scenario correctly, seven participants did not answer the filler question for the headlamp scenario correctly, one also failed the match scenario. Thus, forty-four responses were included in the analyses reported below. Thirteen participants failed the control question for the third scenario we used. Because of the high number we concluded that there was a problem with the material used and excluded this scenario from the evaluations.

### 4.4.2 Results & Discussion

The table in Figure 4.2 states the results of the study. We counted both values 3 and 4 on the scale as judging the sentence true and 0 an 1 as judging the sentence false. The graph in Figure 4.2 plots the percentages of the different answers first for both scenario's separately and then combined. The results show, first of all, that a majority of the participants agree with the intuitions reported by Fisher Fisher [2017]. Furthermore, the results for the opposite counterfactuals (12a) and (12b) support the conclusion that the judgements are for the most part judgements about truth values and not well-formedness of the sentences under consideration.

Hence, we conclude with Fisher that these nested counterfactuals present a problem for the interventionist approach to their meaning. Fisher discussed the possibility to defend the approach by arguing that the conditionals under discussion are interpreted according to a different (epistemic) reading of counterfactuals and eventually dismisses it. We agree with Fisher. Notice the particularity of the situation. Normally, the possibility of an epistemic reading is considered in case a counterfactual intuitively appears to be true, but the account under discussion cannot predict this.[7] Here we would have to explain

[7] A good example are backtracking counterfactuals: counterfactuals reasoning backward in time. The interventionist approach predicts all backtracking to be impossible. However, sometimes backtracking seems to be possible. This is occasionally explained by discussing a

| Sentence | True | False |
|---|---|---|
| (10) | 4% | 80% |
| (12a) | 64% | 13% |
| (11) | 7% | 84% |
| (12b) | 77% | 9% |
| (10)+(11) | 6% | 82% |
| (12a)+(12b) | 71% | 11% |

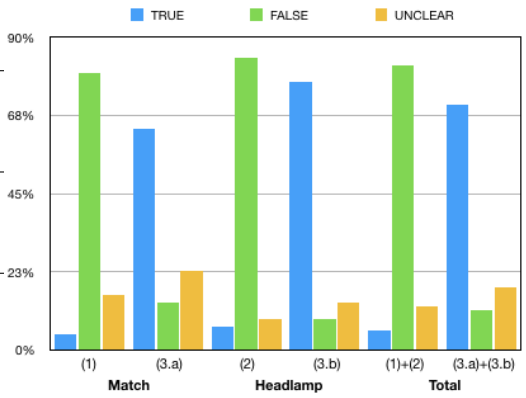



Figure 4.2: Results of the $1^{st}$ study.

why certain counterfactuals are intuitively false, while the approach predicts them to be true. In order to make this work, we would first have to argue that an intervention-based reading of these particular counterfactuals is not possible.

To sum up, the results of this study support Fisher's argumentation against the interventionist approach. But does that mean that we need to give up the interventionist approach to counterfactuals? We don't think so. We can give up the property of strict interventionism responsible for the problematic predictions, but still keep the general idea and all the strong predictions of the interventionist approach. The big conceptual step that needs to be taken is to apply intervention to the valuation $\mathcal{A}$ instead of the representation of the causal dependencies $\mathcal{S}$. In the next section we develop this idea in detail.

## 4.5 Non-strict interventions

The goal is, then, to find a notion of intervention that coincides with Pearl [2000]'s proposal for non-nested cases (thus 'inheriting' the good behaviour of the strict interventionism approach in those situations), but also accounts for the results of the empirical study. The idea on which we will build our alternative proposal is that, although counterfactual assumptions might modify the value of some causal variables, they will not affect causal relationships. This way, the framework can satisfy Fisher's adequacy condition for approaches to the meaning of counterfactuals (see p. 74): even after intervention on a particular variable, it remains connected to its causal parents.

If an intervention can change assignments without modifying causal relationships, we might end up with models in which the values of variables (as defined by the assignment) do not comply with the laws (as defined by the

possible epistemic reading that allows for backtracking.

causal dependencies).[8] The notion of a causal model introduced in Section 4.2 doesn't allow for this. We need a more general notion of causal model, one that does not require the assignment $\mathcal{A}$ to comply with the structural functions in $\mathcal{S}$.

**23.** Definition (General causal model). *A* general *causal model over $\mathcal{U} \cup \mathcal{V}$ is a tuple $\langle \mathcal{S}, \mathcal{A} \rangle$ in which $\mathcal{S}$ is defined as in Definition 18 and $\mathcal{A}$ is defined as a function assigning to every $X \in (\mathcal{U} \cup \mathcal{V})$ a value $\mathcal{A}(X) \in \mathcal{R}(X)$.*

With this generalized notion of a causal model at hand, we can now introduce a new notion of intervention that captures the idea described above: it modifies the value of causal variables, but leaves causal relationships unaffected. This new mode of intervention will be expressed by a different type of sentence in the formal language we use.

**24.** Definition. *Formulas $\phi$ of the language $\mathcal{L}_{[\,],\boxdot\!\!\rightarrow}$ over $\mathcal{U} \cup \mathcal{V}$ are given by*

$$\phi ::= X{=}x \mid \neg\phi \mid \phi \wedge \phi \mid [\overrightarrow{X}{=}\overrightarrow{x}]\phi \mid (\overrightarrow{X}{=}\overrightarrow{x}) \boxdot\!\!\rightarrow \phi$$

*for $X \in \mathcal{U} \cup \mathcal{V}$, $x \in \mathcal{R}(X)$, $\overrightarrow{X} = (X_1, \ldots, X_k) \in (\mathcal{U} \cup \mathcal{V})^k$, $k \in \mathbb{N}$, $X_i \neq X_j$ for $i \neq j$, $\overrightarrow{x} = (x_1, \ldots, x_k)$ with $x_i \in \mathcal{R}(X_i)$.*

We have now two counterfactual formulas: $[\overrightarrow{X}{=}\overrightarrow{x}]\phi$ and $(\overrightarrow{X}{=}\overrightarrow{x}) \boxdot\!\!\rightarrow \phi$. The former will be semantically interpreted using the well-known notion of intervention described in Definition 22. We will refer to this notion of intervention as *strict* intervention. The latter will be semantically interpreted using our new notion of *non-strict* intervention. This notion captures the following intuition: a non-strict intervention affects only the value of the variables that are causally dependent on the intervened ones; their values should be set according to the causal laws. The rest of the variables, those causally independent of the intervened ones, should remain untouched. This idea is formally spelled out in Definition 25.

**25.** Definition. *Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a general causal model. When evaluating formulas in $\mathcal{L}_{[\,],\boxdot\!\!\rightarrow}$, the semantic interpretation of formulas also in $\mathcal{L}_{[\,]}$ is as in Definition 22. For formulas of the form $(\overrightarrow{X}{=}\overrightarrow{x}) \boxdot\!\!\rightarrow \phi$,*

$$\langle \mathcal{S}, \mathcal{A} \rangle \models (\overrightarrow{X}{=}\overrightarrow{x}) \boxdot\!\!\rightarrow \phi \qquad \textit{iff}_{def} \qquad \langle \mathcal{S}, \mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}} \rangle \models \phi$$

*with $\langle \mathcal{S}, \mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}} \rangle$ the general causal model whose assignment, $\mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}}$, is obtained in the following way. Let $\overrightarrow{X_d}$ be a vector containing the variables in $\overrightarrow{X}$ whose current value (as given by $\mathcal{A}$) is different from their intended new value (as indicated by $\overrightarrow{x}$).*

[8] Namely, there might be variables whose values is different from the one obtained by using their structural function with the values of all other variables.

**(i)** *The value of variables in* $\overrightarrow{X}$ *becomes* $\overrightarrow{x}$ *(as indicated by the intervention).*

**(ii)** *For each variable* $Y$ *not in* $\overrightarrow{X}$*,*

(a) *if* $Y$ *is not causally dependent on any variable in* $\overrightarrow{X_d}$ *(i.e., if there is no* $X_d \in \overrightarrow{X_d}$ *such that* $X_d \hookrightarrow^+_{\mathcal{S}} Y$*), keep its value as in* $\mathcal{A}$*.*

(b) *if* $Y$ *is causally dependent on some variables in* $\overrightarrow{X_d}$ *(i.e., if* $X_d \hookrightarrow^+_{\mathcal{S}} Y$ *for some* $X_d \in \overrightarrow{X_d}$*), its value is calculated according to the causal laws in* $\mathcal{S}$ *from the values already in* $\mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}}$*.*[9]

The notions of strict (Definition 22) and non-strict intervention (Definition 25) differ in two crucial points. The first difference is in the structural functions of the resulting models. In the model resulting from a strict intervention, the intervened variables have been cut off from their causal parents; however, the model resulting from the just defined non-strict intervention preserves the previous causal information.[10] The second difference concerns the way the new assignment is defined for endogenous variables. In the strict interventionist case, the values of *all* endogenous variables are recalculated according to the (recall: modified) structural functions. In the non-strict case, recalculation (recall: with respect to the original structural functions) takes place only for endogenous variables causally dependent on the variables intervened on.[11]

The following observation will be useful: in a model where the valuation complies with the structural functions, the model that results from a non-strict intervention can be equivalently defined in the following way.

**4.5.1.** Proposition. *Let* $\langle \mathcal{S}, \mathcal{A} \rangle$ *be a causal model (i.e., a general causal model where* $\mathcal{A}$ *complies with* $\mathcal{S}$*); let* $\overrightarrow{X}=\overrightarrow{x}$ *be the antecedent of a counterfactual formula. Let*

- $\overrightarrow{X_d}$ *be as before: the variables in* $\overrightarrow{X}$ *whose value (as given by* $\mathcal{A}$*) differs from their intended new value (as indicated by* $\overrightarrow{x}$*);*
- $\overrightarrow{Y}$ *be the endogenous variables not occurring in* $\overrightarrow{X}$ *that are not causally dependent on variables in* $\overrightarrow{X_d}$*, with* $\overrightarrow{y}$ *their values according to* $\mathcal{A}$*.*

[9] Recall: we work only with recursive models. Thus, from $\mathcal{S}$'s induced causal graph $\langle \mathcal{U} \cup \mathcal{V}, \hookrightarrow \rangle$ and the antecedent $\overrightarrow{X}=\overrightarrow{x}$, one can create a chain of sets of variables $S_0 \subseteq \cdots \subseteq S_n$ where $S_0 = \overrightarrow{X} \cup (\mathcal{U} \setminus \overrightarrow{X})$, $S_n = \mathcal{U} \cup \mathcal{V}$ and, for any $S_i$ and $S_{i+1}$, the value of variables in $S_{i+1} \setminus S_i$ can be calculated from the causal dependencies and the value of variables in $S_i$. The values the valuation $\mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}}$ assigns to variables in $S_0$ are fixed from the initial valuation $\mathcal{A}$ and the antecedent $\overrightarrow{X}=\overrightarrow{x}$, so the values of the rest can be properly obtained.

[10] Thus, in the latter, the valuation $\mathcal{A}^{\overrightarrow{X}=\overrightarrow{x}}$ may not comply with $\mathcal{S}$'s structural functions.

[11] Note: when the original assignment $\mathcal{A}$ complies with the causal dependencies in $\mathcal{S}$, both strategies produce the same result.

*Then, the assignment $\mathcal{A}^{\vec{X}=\vec{x}}$ (Definition 25) can be equivalently defined as the (unique) assignment that is identical with $\mathcal{A}$ with respect to exogenous variables not in $\vec{X}$, assigns to exogenous variables in $\vec{X}$ their respective value in $\vec{x}$, and complies with the causal dependencies in $\mathcal{S}_{(\vec{X}=\vec{x},\vec{Y}=\vec{y})}$ (see Definition 22).*[12]

If $\langle \mathcal{S}, \mathcal{A} \rangle$ is a model without causal violations (i.e., $\mathcal{A}$ complies with $\mathcal{S}$), then the assignment created by a non-strict intervention ($\mathcal{A}^{\vec{V}=\vec{v}}$) coincides with the one created by a strict intervention ($\mathcal{A}^{\mathcal{S}_{\vec{V}=\vec{v}}}$).

### 4.5.1 Fisher's counter-examples revisited

The semantics for counterfactuals proposed here can deal with the examples **Match** and **Headlamp** discussed in Section 4.3 and 4.4. For reasons of space we will only discuss **Match** (**Headlamp** works analogously).

- **Match.** I hold up a match and strike it, but it does not light. I say

  (13) If the match had lit, then (even) if it had not been struck, it would have lit.

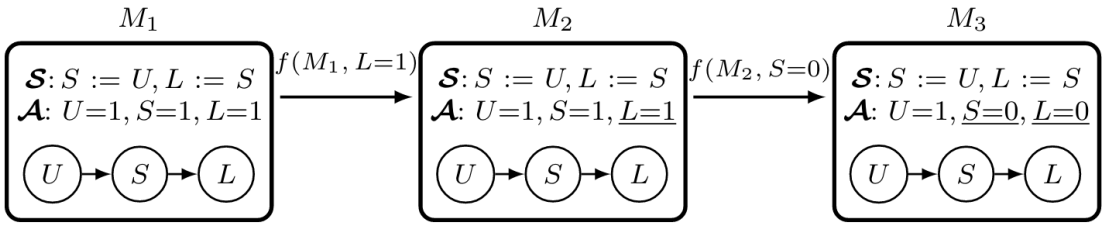



Figure 4.3: The evaluation of the **Match** example with the selection function $f$

First, we need to define the causal model $M_1 = \langle \mathcal{S}, \mathcal{A} \rangle$ with respect to which the counterfactual (13) is interpreted. We define $\mathcal{V} = \{S, L\}$ and $\mathcal{U} = \{U\}$, with $S$ indicating whether the match has been struck (1:yes, 0:no) and $L$ indicating whether the match has lit (1:yes, 0:no). The exogenous variable $U$ represents external factors causally responsible for $S$.[13]

The sentence contains nested counterfactuals, so we need to intervene twice: first, with $L{=}1$ (the antecedent of the main counterfactual), and then, with $S{=}0$ (the antecedent of the embedded counterfactual). On the resulting model, we should check whether $L{=}1$ (the consequent of the embedded counterfactual) is

[12] The proof of this proposition can be found in the appendix of Chapter 4

[13] Note: our setting allows intervention on exogenous variables, so $S$ can be take to be exogenous, thus making $U$ superfluous. Still, $U$ is kept, in line with the common modelling strategy of representing external factors by means of exogenous variables.

true. The first intervention, $L$=1, produces model $M_2$ in Figure 4.3 (Definition 25), affecting the original assignment but preserving the original causal dependencies. For evaluating the embedded counterfactual $(S{=}0) \boxright L = 1$, we apply the second intervention, $S$=0, to $M_2$. This results in the model $M_3$ in Figure 4.3, with $S = 0$ as the intervention requires, and $L = 0$, as $L$'s value is still causally sensitive to $S$. In this final model, the innermost consequent $L = 1$ fails; thus,

$$M_1 \not\models (L{=}1) \boxright ((S{=}0) \boxright L = 1).$$

We correctly predict that the counterfactual (13) is false in the given context.

### 4.5.2 The Import/Export Principle

There is an interesting connection between the core examples discussed in this paper and the famous Import-Export Principle (IEP): if $A$ and $B$ share none of their variables, then $(A \wedge B) \boxright C$ is equivalent to $A \boxright (B \boxright C)$.[14] This principle got a lot of attention in the literature on the logic of conditional sentences, in particular in connection to indicative conditionals (Lewis [1973] , Skyrms [1980] among many others). But while for indicative conditionals it is generally accepted that the principle should be valid, this is less clear for counterfactuals. Because the similarity approach predicts the principle to fail[15], people have been looking for examples confirming this prediction(Etlin [2008], Kaufmann [2005], Starr [2019]). Interestingly, these examples are often identical or very similar to the core examples of Fisher that we are trying to account for here. A rare exception is (14a), brought forward in Skyrms [1980]. Skyrms considers this sentence to be true and the related counterfactual (14b) where both antecedent of (14a) are combined in one antecedent false.

(14) a. If this sample were burning green (say it was barium) then it would still be true that had it been sodium it would have burned yellow.

b. If the sample were burning green and had been sodium, it would have burn yellow.

Another example can be found in Lange [1999].

> Suppose that you and I have just run a race, and I have won. I believe that I would always win if I really tried. Then I am willing to assert: "Suppose that you had won the race. Then I must not have

[14] I.e. $((A \wedge B) \boxright C) \leftrightarrow (A \boxright (B \boxright C)$ is valid. This principle is sometimes also called the Weak Import-Export Principle, while the principle without the restriction of non-common variables is called Import-Export Principle.

[15] But see Starr [2014] for a dynamic semantic implementation of the similarity analysis that does validate the Import-Export Principle.

> been trying; had I tried, I would have won." This is $p > (q > r)$. I am not willing to assert the corresponding $(p \wedge q) > r$: Had you won and I really tried, I would have won. There is no logically possible world in which you and I both win the race. (Lange [1999], p.259)

However, all the examples brought forward as violations of IEP share the causal structure of the core examples discussed in our paper: IEP is observed to fail in case the first antecedent in a right-nested counterfactual causally depends on the second, embedded antecedent. In this situation the second antecedent can overwrite the truth of the first antecedent. But if the counterfactual is reformulated in conjunctive form this overruling of the first antecedent is not possible anymore.

While the similarity approach does predict that IEP fails, it cannot explain the specific circumstances in which the principle breaks down and why it seems to hold in so many other cases. Now, one would expect an interventionist account to outperform the similarity approach here, because of the central role causality plays in the interventionist picture. But a strict interventionalist approach like the one introduced in Section 4.2 validates IEP.[16] Recall that in strict interventionism the variables that are intervened on are cut off from their causal parents and forced to a particular value. Then all endogenous variables are calculated again from the values of the exogenous variables and the new causal dependencies. The order in which interventions are executed has no effect on the result: once a variable is intervened on, no later intervention on different variables can change its value. So, the particular observation that we are considering here, where a later intervention overrules an earlier one, cannot be modeled.

Our account, however, is made exactly to deal with these causal exceptions to the IEP (see Section 4.5.1). We allow for later interventions to overwrite the effect of earlier ones. This occurs exactly in the case the later intervention affects causes of the earlier intervention. If the antecedents of a right nested counterfactual are causally independent of each other, our approach will predict that IEP is valid.[17] But in general IEP is not valid. Our approach improves on the similarity approach, because it puts the finger much more precisely on the point were IEP fails.

[16] For discussion and a proof see Briggs [2012].

[17] In this case the order in which the interventions are performed has no effect on the resulting model. This follows immediately from the fact that an intervention, as the notion is defined here, will only affect the value of the variables intervened on and variables causally dependent on these variables.

| | | |
|---|---|---|
| **A0** | Propositional tautologies | **MP** From $\phi$ and $\phi \to \psi$ infer $\psi$ |
| **A1** | $[\vec{X}{=}\vec{x}](Y = y) \to \neg[\vec{X}{=}\vec{x}](Y = y')$ | for $y, y' \in \mathcal{R}(Y)$ with $y \neq y'$ |
| **A2** | $\bigvee_{y \in \mathcal{R}(Y)} [\vec{X}{=}\vec{x}](Y = y)$ | |
| **A3** | $\big([\vec{X}{=}\vec{x}](Y = y) \wedge [\vec{X}{=}\vec{x}](Z = z)\big) \to [\vec{X}{=}\vec{x}, Y{=}y](Z = z)$ | |
| **A4** | $[\vec{X}{=}\vec{x}, Y{=}y](Y = y)$ | |
| **A5** | $\big([\vec{X}{=}\vec{x}, Y{=}y](Z = z) \wedge [\vec{X}{=}\vec{x}, Z{=}z](Y = y)\big) \to [\vec{X}{=}\vec{x}](Z = z)$ | for $Y \neq Z$ |
| **A6** | $(X_0 \rightsquigarrow X_1 \wedge \cdots \wedge X_{k-1} \rightsquigarrow X_k) \to \neg(X_k \rightsquigarrow X_0)$ | |
| **A7** | $[\vec{X}{=}\vec{x}](\phi \wedge \psi) \leftrightarrow ([\vec{X}{=}\vec{x}]\phi \wedge [\vec{X}{=}\vec{x}]\psi)$ | |
| **A8** | $[\vec{X}{=}\vec{x}]\neg\phi \leftrightarrow \neg[\vec{X}{=}\vec{x}]\phi$ | |
| **A9** | $[\vec{X}{=}\vec{x}][\vec{Y}{=}\vec{y}]\psi \leftrightarrow [\vec{X'}{=}\vec{x'}][\vec{Y}{=}\vec{y}]\psi$ | for $\vec{X'} = \vec{X} \setminus \vec{Y}$ |
| **A10** | $(Y = y) \to \neg(Y = y')$ | for $y, y' \in \mathcal{R}(Y)$ and $y \neq y'$ |
| **A11** | $\bigvee_{y \in \mathcal{R}(Y)} (Y = y)$ | |
| **A12** | $(\vec{X} = \vec{x}) \boxdot\!\!\to \bigvee_{y \in \mathcal{R}(Y)} (Y = y)$ | |
| **A13** | $\bigwedge \left\{ \begin{array}{l} \bigwedge_{X_i \in \vec{X} \cap \vec{X_d}} (X_i \neq x_i), \\ \bigwedge_{X_i \in \vec{X} \setminus \vec{X_d}} (X_i = x_i), \\ \bigwedge_{Z \in \vec{Z}} \neg \bigvee_{X_d \in \vec{X_d}} (X_d \rightsquigarrow^{+} Z), \\ \bigwedge_{Z' \in \mathcal{V} \setminus \vec{Z}} \bigvee_{X_d \in \vec{X_d}} (X_d \rightsquigarrow^{+} Z') \end{array} \right\} \to \big((\vec{X} = \vec{x}) \boxdot\!\!\to Y = y) \leftrightarrow [\vec{X}{=}\vec{x}, \vec{Z}{=}\vec{z}]Y = y\big)$ | |
| **A14** | $((\vec{X}{=}\vec{x}) \boxdot\!\!\to [\vec{Y}{=}\vec{y}]\phi) \leftrightarrow ([\vec{X'}{=}\vec{x'}][\vec{Y}{=}\vec{y}]\phi)$ | for $\vec{X'} = \vec{X} \setminus \mathcal{V}$ |
| **A15** | $\big((\vec{X}{=}\vec{x}) \boxdot\!\!\to (\phi \wedge \psi)\big) \leftrightarrow \big((\vec{X}{=}\vec{x}) \boxdot\!\!\to \phi \ \wedge \ (\vec{X} = \vec{x}) \boxdot\!\!\to \psi\big)$ | |
| **A16** | $((\vec{X} = \vec{x}) \boxdot\!\!\to \neg\phi) \leftrightarrow \neg((\vec{X} = \vec{x}) \boxdot\!\!\to \phi)$ | |

Table 4.1: Axiom system for $\mathcal{L}_{[\,],\,\boxdot\!\!\to}$ w.r.t. causal models.

### 4.5.3 The Axiomatization for the logic

Note how direct dependency (Definition 19) is syntactically definable in terms of strict intervention. Recall that $V \in \mathcal{V}$ is directly dependent on $X \in \mathcal{U} \cup \mathcal{V}$ (notation: $X \hookrightarrow V$) if and only if there is an assignment $\vec{z}$ of values for all variables in $\vec{Z} = \mathcal{U} \cup \mathcal{V} \setminus \{X, V\}$ and two different values $x_1, x_2$ of $X$ such that the value $V$ gets by setting $(\vec{Z}, X)$ to $(\vec{z}, x_1)$ is different from the value it gets by setting the same variables to $(\vec{z}, x_2)$. This can be expressed by the formula

$$\bigvee_{\substack{\vec{z} \in \mathcal{R}(\mathcal{U} \cup \mathcal{V} \setminus \{X, V\}), \\ \{x_1, x_2\} \subseteq \mathcal{R}(X), x_1 \neq x_2, \\ \{v_1, v_2\} \in \mathcal{R}(V), v_1 \neq v_2}} [\vec{Z}{=}\vec{z}, X{=}x_1](V = v_1) \ \wedge \ [\vec{Z}{=}\vec{z}, X{=}x_2](V = v_2),$$

which will be abbreviated as $X \rightsquigarrow V$ (cf. with the syntactic definition of causal dependency in Halpern [2000]). Moreover: thanks to the finiteness of the sets

of variables, the notion of causal dependency (the transitive closure of direct dependency, $\hookrightarrow^+$) is also syntactically definable. Indeed, given that $|\mathcal{U}| = m$ and $|\mathcal{V}| = n$, the fact that $V$ is causally dependent on $X$ can expressed by the formula

$$(X \rightsquigarrow V) \vee \bigvee_{k=1}^{m+n-2} \bigvee_{\{\langle X_1,\ldots,X_k\rangle | X_i \in (\mathcal{U}\cup\mathcal{V}\setminus\{X,V\})\}} \Big((X \rightsquigarrow X_1) \wedge \bigwedge_{j=1}^{k-1} (X_j \rightsquigarrow X_{j+1}) \wedge (X_k \rightsquigarrow V)\Big)$$

which will be abbreviated as $X \rightsquigarrow^+ V$.[18]

With this syntactic abbreviation, and thanks to Proposition 4.5.1, it is possible to axiomatise the modified notion of intervention, with the system presented in Table 4.1. The first block deals with propositional validities. The second axiomatises the strict intervention [ ] taking advantage of the fact that our proposal is a conservative extension of the original causal modelling semantics Pearl [2000].[19] Axioms **A1**-**A8** characterise its basic (*non-nested*) behaviour Halpern [2000]. Then, axiom **A9** specifies the way a nested strict intervention works Briggs [2012], **?**: if a model is strictly intervened first on $\overrightarrow{X}$ and on the disjoint set of variables $\overrightarrow{Y}$, then the former intervention will be overwritten by the latter.

The third block characterise the behaviour of our non-strict intervention $\boxdot\!\!\rightarrow$. Axioms **A10**-**A11** indicate that every variable has exactly one value.[20] Then, axiom **A12** states that our modified version of intervention assigns proper values. Axioms **A13** and **A14** are the crucial ones, as they describe the relationship between the two forms of intervention. Axiom **A13** relies on Proposition 4.5.1 to describe the assignment after a non-strict intervention $\boxdot\!\!\rightarrow$ in terms of the assignment after a (different) strict intervention [ ]. It states that, if $\overrightarrow{X_d}$ contains exactly the variables in $\overrightarrow{X}$ whose value would change (conjuncts 1 and 2 in the antecedent), and $\overrightarrow{Z}$ contains exactly the variables that are not causally dependent on those in $\overrightarrow{X_d}$ (conjuncts 3 and 4 in the antecedent), then a non-strict-intervention with $\overrightarrow{X}=\overrightarrow{x}$ coincides with a strict intervention with $\overrightarrow{X}=\overrightarrow{x}, \overrightarrow{Z}=\overrightarrow{z}$. Axiom **A14** states that non-strict intervention on endogenous variables does not affect the truth of formulas within the scope of strict intervention, since causal relationships are invariant under non-strict interventions.

[18] In the formula, $k$ runs over the number of needed intermediate variables (at least 1, and at most $m + n - 2$). Then, the intermediate disjunct runs over all possible tuples of $k$ variables.

[19] The semantic interpretation of atoms, Boolean operators and the strict intervention are as in Pearl [2000]. The class of models do changes, but this affects neither atoms (variables still have exactly one of their allowed values) nor Boolean operators. Crucially, strict-interventions force the assignment to agree with the structural functions, so formulas occurring under their scope behave just as under the original causal models.

[20] In Halpern [2000], $X = x$ is equivalent to [ ]$(X = x)$; thus, axioms **A1** and **A2** suffice. This is not the case in our setting, as interventions cannot be empty; hence the need of **A10**-**A11**.

Finally, axioms **A15-A16** are the rules for Boolean operators: given a certain causal model, there is exactly one causal model results from a certain non-strict intervention, so Boolean operators can be distributed or pushed into $\Box\!\!\rightarrow$.

**4.5.2.** THEOREM. *This axiom system is sound and strongly complete with respect to recursive causal models.* [21]

## 4.6 Discussion and conclusions

In this paper we proposed a new approach to the semantics of counterfactual conditionals. Our proposal builds on the well-known interventionist approach, but uses a different approach to intervention. There are two separate steps that we took in defining our proposal. First, we made a substantial conceptual shift in what we understand to be the target of intervention. We propose that intervention does not take place at the level of structural dependencies, but at the level of the (incidental) valuations of the variables. Conceptually, this means that we see intervention not as a hypothetical modification of the underlying laws of nature, but as the hypothetical assumption of exceptions to the laws (see Schulz [2011, 2014] for a similar move). As a consequence, after intervention, no information on causal dependencies in the actual world is lost. The second part of the proposal lies in how exactly we define the valuation resulting from intervention. We propose, on the one hand, that the value of variables not causally affected by the intervened variables remains unchanged and, on the other hand, that the value of the causally dependent variables is recalculated according to the laws, the new value of the intervened variables and the old values of the causally independent variables (see Definition 25). This approach allows us to satisfy our objectives: (i) the predictions made for the truth conditions of counterfactuals that are not right-nested are the same as made in Briggs [2012], and (ii) the approach correctly deals with the counterexamples brought forward in Fisher [2017].

The change we propose for the concept of intervention, though minor in terms of predictions, is conceptually quite substantial. In future work we hope to provide more evidence showing that such a radical change is needed. For instance, we should look for other counterfactuals for which both notions of intervention make different predictions, and then test which approach better matches the intuitions of speakers. Remember that Fisher only discusses examples of the form (i) $B \Box\!\!\rightarrow (\neg A \Box\!\!\rightarrow B)$, where $A$ is a cause of $B$, while he claims that the observation extends to arbitrary right-nested counterfactuals. One way to test our approach would be to look at other types of right-nested counterfactuals, for instance examples of the form (iii) $B \Box\!\!\rightarrow (\neg A \Box\!\!\rightarrow C)$, where $C$ is a direct cause of $B$. In a scenario where $A$ causes $B$ and $B$ causes $C$, if in the actual

[21] Proofs can be found in the appendix of Chapter 4.

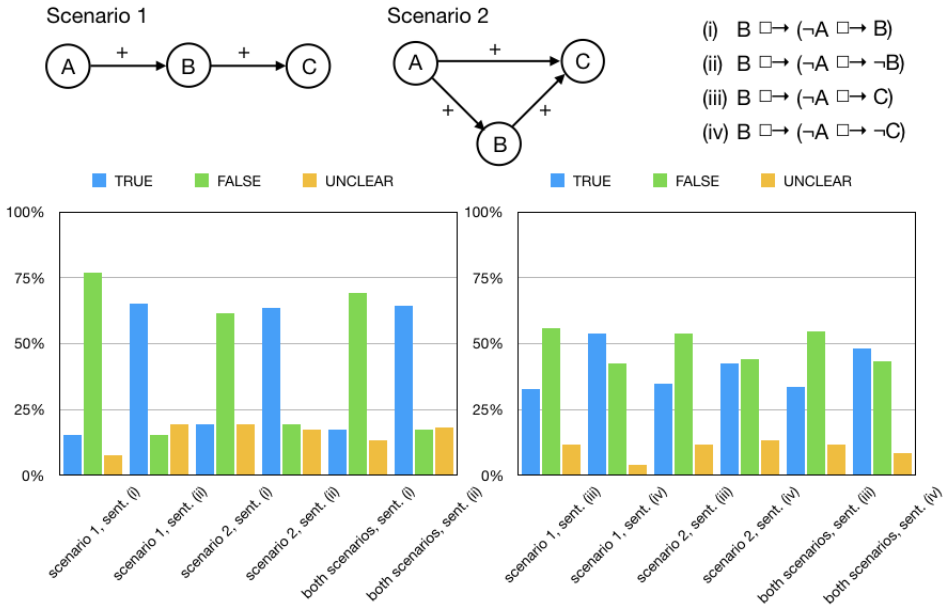



Figure 4.4: Overview of the results of the second study; the sentences (i)-(iv) are those that we asked participants to judge in the two scenarios.

context $A = 1, B = C = 0$ the strong interventionist approach would predict that both (i) and (iii) should be true, while according to our approach these counterfactuals should be false. We performed a second preliminary study to test these predictions (see the two scenarios in Figure 4.4). While we could confirm, using the same method as before[22], that still the majority of the participants consider counterfactual of the form (i) false (left diagram in Figure 4.4), this effect becomes weaker for counterfactuals of type (iii) and basically disappears in combination with scenario 2 (right diagram in Figure 4.4). Notice, that additional variables don't mean a simple increase in uncertainty in the given answers. People still feel that they have intuitions about the truth values of these sentences; it's just that their opinions differ.

These results are problematic for strict interventionism as well as the alternative we proposed here. But one should be careful with over-interpreting the experiments we report on here. Future work will have to show whether the results obtained are stable. And only when we have a clear picture of the phenomenon that we need to account for does it make sense continue modifying the approach. Still we think that our proposal makes an important step in the right direction. Fisher's examples clearly show that sometimes we need to be able to recall causal dependencies after an intervention has violated them. This means that the structural information about these dependencies should not be

[22]The scripts and data are available at http://projects.illc.uva.nl/cil/.

the locus of the intervention. So, what we certainly want to defend here is the proposed step from intervention on the causal dependencies to intervention on the valuation of the variables. Whether the exact form we gave to intervention on the valuation is correct needs to be studied in future work.

## 4.7 Appendix

**Proof of Proposition 4.5.1.** Let $\langle \mathcal{S}, \mathcal{A} \rangle$ be a causal model. Let $\vec{X}{=}\vec{x}$ be a counterfactual antecedent, with ***(i)*** $\vec{X}_d$ the variables in $\vec{X}$ whose $\mathcal{A}$-value differs from the intended $\vec{x}$, and ***(ii)*** $\vec{Y}$ the endogenous variables not occurring in $\vec{X}$ that are not causally dependent on variables in $\vec{X}_d$, with $\vec{y}$ their $\mathcal{A}$-values.[23] Abbreviate $\mathcal{A}^{\vec{X}=\vec{x}}$ (Definition 25) as $\mathcal{A}_1$, and let $\mathcal{A}_2$ be the (unique) assignment that is identical with $\mathcal{A}$ with respect to exogenous variables not in $\vec{X}$, assign to exogenous variables in $\vec{X}$ their respective value in $\vec{x}$, and complies with the structural functions in $\mathcal{S}_{(\vec{X}=\vec{x},\vec{Y}=\vec{y})}$. Take any $Z \in \mathcal{U} \cup \mathcal{V}$; it will be shown that $\mathcal{A}_1(Z) = \mathcal{A}_2(Z)$.

First, for *exogenous* variables. ***(1)*** If $Z$ is not in $\vec{X}$, then both $\mathcal{A}_1$ and $\mathcal{A}_2$ agree with $\mathcal{A}$, the first because $Z$ is not causally dependent on any other variable, and the second by definition. ***(2)*** If $Z$ is in $\vec{X}$, both $\mathcal{A}_1$ and $\mathcal{A}_2$ give it the value indicated by $\vec{x}$.

Now, for *endogenous* variables. ***(1)*** If $Z$ is in $\vec{X}$, then $\mathcal{A}_1$ gives it the value indicated by $\vec{x}$ (by definition), and so does $\mathcal{A}_2$ (as it complies with $\mathcal{S}_{(\vec{X}=\vec{x},\vec{Y}=\vec{y})}$). ***(2)*** If $Z$ is in $\vec{Y}$ then, $\mathcal{A}_1$ uses the value in $\mathcal{A}$ (by definition), and so does $\mathcal{A}_2$ (via the values $\vec{y}$, taken from $\mathcal{A}$). ***(3)*** Finally, suppose $Z$ is neither in $\vec{X}$ nor in $\vec{Y}$; then, the structural functions used by $\mathcal{A}_1$ and $\mathcal{A}_2$ to calculate $Z$'s value are the same: from $\mathcal{S}$ for the first, and from $\mathcal{S}_{(\vec{X}=\vec{x},\vec{Y}=\vec{y})}$ for the second). On its own, this does not guarantee that $Z$'s *value* under both $\langle \mathcal{S}, \mathcal{A}_1 \rangle$ and $\langle \mathcal{S}_{(\vec{X}=\vec{x},\vec{Y}=\vec{y})}, \mathcal{A}_2 \rangle$ is the same: there is a unique $f_Z$, and yet the values of its parameters (all other variables) might be different. But, according to the previous items, the only variables in which $\mathcal{A}_1$ and $\mathcal{A}_2$ might differ are precisely the endogenous variables in neither $\vec{X}$ nor in $\vec{Y}$. Then, relying on the *recursiveness* of the model, one can use an inductive argument to show that, when the process that assigns values to variables calculates the value of such a $Z$, the values of all its *parents* will be the same in both $\mathcal{A}_1$ and $\mathcal{A}_2$. The step $\#_0$ in the process assigns values to the variables in the set $S_0 := \{Z \in \mathcal{V} \backslash (\vec{X} \cup \vec{Y}) \mid$ the parents of $Z$ are in $\mathcal{U} \cup \vec{X} \cup \vec{Y}\}$. The valuations $\mathcal{A}_1$ and $A_2$ coincide in the values of the variables in $\mathcal{U} \cup \vec{X} \cup \vec{Y}$,

[23] Thus, $\vec{X}$ and $\vec{Y}$ are disjoint.

so they will coincide in the values of variables in $S_0$. Crucially, the model is recursive, so $S_0 \neq \emptyset$. Then, each step $\#_{k+1}$ assigns values to the variables in the set $S_{k+1} := \{Z \in \mathcal{V} \setminus (\vec{X} \cup \vec{Y}) \mid \text{the parents of } Z \text{ are in } \mathcal{U} \cup \vec{X} \cup \vec{Y} \cup \bigcup_{0 \leq i \leq k} S_i\}$. Now, $\mathcal{A}_1$ and $A_2$ coincide in the values of the variables in $\mathcal{U} \cup \vec{X} \cup \vec{Y} \cup \bigcup_{0 \leq i \leq k} S_i$, so they will coincide in the values of variables in $S_{k+1}$. Crucially again, the model is recursive, so $S_{k+1} \neq \emptyset$. Thus, eventually the values of all variables in $\mathcal{V} \setminus (\vec{X} \cup \vec{Y})$ will be calculated, and the values will be the same in both $A_1$ and $A_2$.

**Sketch of proof for Theorem 4.5.2.** The proof follows the *reduction* axioms strategy frequently used in *dynamic epistemic logic* Baltag et al. [1998], van Ditmarsch et al. [2008], van Benthem [2011]. In our case, the strategy relies on a sound and complete axiom system for the $\Box\!\!\rightarrow$-less fragment of $\mathcal{L}_{[],\Box\!\!\rightarrow}$. For the remaining formulas, those involving $\Box\!\!\rightarrow$, the strategy uses 'reduction axioms': valid formulas and validity-preserving rules indicating how to translate a formula with occurrences of $\Box\!\!\rightarrow$ into a provably equivalent one without them. Soundness follows from the validity and validity-preserving properties of the new axioms and rules (so a formula and its translation are semantically equivalent); completeness follows from the completeness of the axiom system for the $\Box\!\!\rightarrow$-less fragment, as the recursion axioms define a recursive validity-preserving translation from the full $\mathcal{L}_{[],\Box\!\!\rightarrow}$ into the latter. The reader is referred to [van Ditmarsch et al., 2008, Chapter 7] for a detailed explanation of this technique.

For the underlying system, propositional calculus and axioms **A1-A11** constitute a sound and complete axiomatization for $\mathcal{L}_{[]}$ over *general* causal models.[24] For dealing with $\Box\!\!\rightarrow$, axiom **A12** states that our modified version of intervention still assigns variables a proper value.

Axioms **A13-A16** define the recursive translation that takes any formula in $\mathcal{L}_{[],\Box\!\!\rightarrow}$ and returns a logically equivalent one without $\Box\!\!\rightarrow$. Axiom **A13** is the basic case for the translation, as it eliminates $\Box\!\!\rightarrow$ by showing how the assignment that results from a non-strict intervention is equivalent to an assignment that results from a strict intervention (its validity follows from Proposition 4.5.1). Axiom **A14** eliminates a non-strict intervention that precedes a strict one.[25] Finally, **A15** and **A16** indicate how to deal with negations (commute $\Box\!\!\rightarrow$

[24] Note: as proved in Halpern [2000], **A1-A9** constitute a sound and complete axiom system for $\mathcal{L}_{[]}$ over causal models (where the assignment *agrees* with causal dependencies). The additional axioms **A10-A12**, clearly sound, make this work also for the cases in which the assignment does not align with causal dependencies, which might occur as a result of our non-strict intervention.

[25] For its validity, take a causal model $\langle \mathcal{S}, \mathcal{A} \rangle$, with $\vec{u}$ the assignment of $\mathcal{A}$ to exogenous variables. ***(i)*** By semantic interpretation, $\langle \mathcal{S}, \mathcal{A} \rangle \models [\vec{X}{=}\vec{x}]\phi$ holds if and only if $\langle \mathcal{S}_{\vec{X}=\vec{x}}, \mathcal{A}^{\mathcal{S}_{\vec{X}=\vec{x}}} \rangle \models \phi$ holds, with $\mathcal{A}^{\mathcal{S}_{\vec{X}=\vec{x}}}$ the unique solution to $\mathcal{S}_{\vec{X}=\vec{x}}$ whose assignment to exogenous variables is $\vec{u}$. ***(ii)*** Also from semantic interpretation, $\langle \mathcal{S}, \mathcal{A} \rangle \models (\vec{V} = \vec{v}) \Box\!\!\rightarrow [\vec{X}{=}\vec{x}]\phi$ holds if and only if

and $\neg$) and conjunctions (distribute $\Box\!\!\rightarrow$ over $\wedge$).[26]

$f(\langle \mathcal{S}, \mathcal{A}\rangle, \overrightarrow{V} = \overrightarrow{v}) = \langle \mathcal{S}, \mathcal{A}'\rangle \models [\overrightarrow{X}{=}\overrightarrow{x}]\phi$, which in turn holds if and only if $\langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}'^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}\rangle \models \phi$, with $\mathcal{A}'^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}$ the unique solution to $\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}$ whose assignment to exogenous variables is $\overrightarrow{u}$. Then, $\langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}\rangle$ and $\langle \mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}, \mathcal{A}'^{\mathcal{S}_{\overrightarrow{X}=\overrightarrow{x}}}\rangle$ are identical, hence satisfying the same formulas.

[26] Their validity comes from the fact that the non-strict intervention is deterministic.

# Chapter 5

# A Causal Analysis of Modal Syllogisms

## 5.1 Introduction

In his Prior Analytics Aritstotle made a distinction between assertoric and modal syllogistics. The crucial difference between the two syllogistics is that only the latter makes use of two different types of predicative relations: accidental versus essential predication. 'Animal' is essentially predicated of 'men", but 'walking' is not. Although both (a) 'Every man walks' and (b) 'Every man is an animal' can be true, it is natural to say that the 'reasons' for their respective truths are different. Sentence (a) is true by accident, just because every actual man happens to (be able to) walk. The sentence (b), on the other hand, is true because manhood necessarily involves being animate. In traditional terms it is said that (b) is true *by definition*, although this notion of 'definition' should not be thought of nominalistically: it is the *real* definition. A natural way to account for accidental predication is to say that a sentence of the form 'Every $S$ is $P$' is true just in case every actual *S-individual* is also a *P-individual*. But how should we account for essential predication? The answer to this question is important for logic, because it is by now generally assumed that (e.g. Malink [2013], Van Rijen [2012], Thom [1991], Vecchio [2016]) that Aristotle's system of modal syllogisms, which is almost impossible to understand from a modern point of view, should be understood in terms of the difference between accidental and essential predication.

In this paper we will argue for a *causal* analysis of essential predication. We will argue that this fits well with Aristotle's analysis of real definition in the Posterior Analytics, and that in this way we can account in a relatively straightforward way for several puzzling aspects of Aritotle's system of modal syllogisms presented in his Prior Analytics.

## 5.2 Standard and Modal Syllogistics

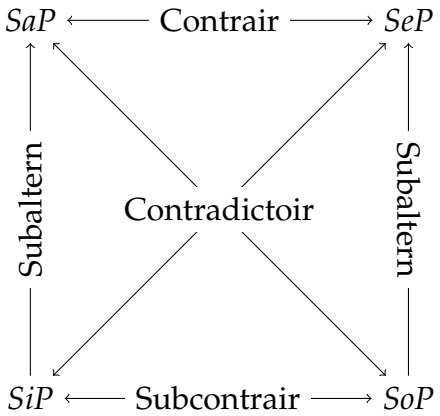
Syllogisms are arguments in which a categorical sentence is derived as conclusion from two categorical sentences as premisses. A categorical sentence is always one of four kinds:

***(i)*** *a*-type: Universal and affirmative ('All men are mortal')

***(ii)*** *i*-type: Particular and affirmative ('Some men are philosophers')

***(iii)*** *e*-type: Universal and negative ('No philosophers are rich')

***(iv)*** *o*-type: Particular and negative ('Some men are not philosophers').




A categorical sentence always contains two *terms*. In the *a*-sentence, for instance, the terms are 'men' and 'mortal', while in the *e*-sentence they are 'philosopher' and 'rich'. Thus, the *syntax* of categorical sentences can be formulated as follows: If *S* and *P* are terms, *SaP*, *SiP*, *SeP*, and *SoP* are categorical sentences. Because a syllogism has two categorical sentences as premisses and one as the conclusion, every syllogism involves only three terms, each of which appears in two of the statements. The first term of the conclusion is called the *subject term*, or *minor term*, the last term, the *predicate term*, or *major term*, and the term that does not occur in the conclusion is called the *middle term*. The premiss in which the major term occurs together with the middle term is called the *major premiss*, the other one the *minor premiss*. The *quality* of a proposition is whether it is *affirmative* (in *a*- and *i*- sentences, the predicate is affirmed of the subject), or *negative* (in *e* and *o*-sentences, the predicate is denied of the subject). Thus 'every man is a mortal' is affirmative, since 'mortal' is affirmed of 'man'. 'No men are immortal' is negative, since 'immortal' is denied of 'man'. The *quantity* of a proposition is whether it is *universal* (in *a*- and *e*-sentences the predicate is affirmed or denied of "the whole" of the subject) or *particular* (in *i* and *o*-sentences, the predicate is affirmed or denied of only 'part of' the subject).

Medieval logicians used the letters '*a*', '*i*', '*e*', and '*o*' for coding the various forms of syllogisms. The *mood* of a syllogism was given by a triple of letters

like *aeo*. This triple, for instance, indicates that the major premiss is of type *a*, the minor premiss of type *e*, and the conclusion of type *o*. But apart from the mood, what is important as well is the *figure*. The figure of a syllogism says whether the major and minor terms occur as subject or predicate in their respective premisses. This gives rise to four possibilities, i.e., four figures:

| 1st | 2nd | 3rd | 4th |
|---|---|---|---|
| MP | PM | MP | PM |
| SM | SM | MS | MS |
| SP | SP | SP | SP |

A *valid* syllogism is a syllogism that cannot lead from true premisses to a false conclusion. It is well-known that by a set theoretic semantic analysis, we can account for syllogistic reasoning. For now we will interpret terms just as sets of individuals and equate for simplicity the interpretation of a term with the term itself. Then we say that $SaP$ is true iff $S \subseteq P$, $SiP$ is true iff $S \cap P \neq \emptyset$, $SeP$ is true iff $S \cap P = \emptyset$, and $SoP$ is true iff $S \not\subseteq P$.[1] This semantic interpretation accounts for many valid syllogisms, but not all of them. In particular, not for the valid syllogisms for which it is required that $SaP$ entails $SiP$. This can be easily accounted for by assuming that for the truth of $SaP$ it is not only required that $S \subseteq P$, but also that $S \neq \emptyset$. It is well-known that with such an interpretation of categorical sentences, all and only all of the following syllogisms are predicted to be valid that Aristotle considered to be valid as well.

| | | | |
|---|---|---|---|
| $\text{Barbara}_1$ | $\text{Barocco}_2$ | $\text{Bocardo}_3$ | $\text{Camenes}_4$ |
| $\text{Celarent}_1$ | $\text{Festino}_2$ | $\text{Disamis}_3$ | $(\text{Fesapo}_4)$ |
| $\text{Darii}_1$ | $\text{Camestres}_2$ | $\text{Ferison}_3$ | $\text{Dimaris}_4$ |
| $\text{Ferio}_1$ | $\text{Cesare}_2$ | $\text{Datisi}_3$ | $\text{Fresison}_4$ |
| $(\text{Barbari}_1)$ | $(\text{Camestrop}_2)$ | $(\text{Felapton}_3)$ | $(\text{Bramantip}_4)$ |
| $(\text{Celaront}_1)$ | $(\text{Cesaro}_2)$ | $(\text{Darapti}_3)$ | $(\text{Camenop}_4)$ |

The syllogisms between brackets are only valid in case one assumes existential import. The above semantic analysis of categorical sentences is nice, because with the help of Venn-diagrams, one can now easily check the validity of any syllogistic argument.[2] For later in the paper, note that we could interpret Aristotle's standard categorical sentences probabilistically as well with equivalent predictions: $SaP$ is true iff the conditional probability of $P$ given $S$ is 1, $P(P|S) = 1$, $SeP$ is true iff $P(S \cap P) = 0$, $SiP$ is true iff $P(S \cap P) \neq 0$ and $SoP$ is true iff $P(P|S) \neq 1$. Notice that on this probabilistic interpretation $SaP$ presupposes that

[1] Warning: in the literature categorical sentences of the form $XaY$ and $XiY$ are read many times in the converse order as we read them and mean that all/some $Y$ belong to $X$.

[2] On the other hand, it is well-known that we don't need the full power of Boolean algebra to account for Syllogistic validity; semi-lattices will do.

$P(S) > 1$, which immediatelly accounts for Aristotle's subalternation inference: $SaP \models SiP$.

Let us now come back to the question what is the natural interpretation of Aristotle's *modal* syllogistics. Let us assume that $Ba^{\Box}C$ means that all $B$s are *necessary*/essentially $C$. Aristotle claims that the following modal syllogisms are valid and invalid, respectively:

| | | | |
|---|---|---|---|
| ***(i)*** | $Ba^{\Box}C, Aa^{\Box}B \therefore Aa^{\Box}C$ | Valid | Barbara LLL |
| ***(ii)*** | $Ba^{\Box}C, AaB \therefore Aa^{\Box}C$ | Valid | Barbara LXL |
| ***(iii)*** | $BaC, Aa^{\Box}B \therefore Aa^{\Box}C$ | Invalid | Barbara XLL |

Although Aristotle had intuitions about which modal syllogistic inferences are valid and which not, he did not base that on a standard semantics. As it turns out, it is already hard enough to account semantically for the intuitions concerning ***(i)***-***(iii)***. But what makes the task especially challenging is that Aristotle also claims that not only conversion inference 4 is valid, but that the same holds for the modal conversion inferences 5 and 6:

| | | |
|---|---|---|
| 4. | $BeC \therefore CeB$ | Valid |
| 5. | $Be^{\Box}C \therefore Ce^{\Box}B$ | Valid |
| 6. | $Bi^{\Box}C \therefore Ci^{\Box}B$ | Valid |

Of course, it is easy to account for inferences 5 and 6 if we assume that the modal should be interpreted in a *de dicto* way. But it is equally easy to see that on such an analysis inference ***(ii)*** is *not* predicted to be valid. A *de re* analysis of sentences like $Ba^{\Box}C$, on the other hand, would make inference 2 valid, but such an analysis cannot account for the modal conversion inferences 5 and 6. So neither a standard *de dicto* nor a *de re* analysis of modal statements would work to account for Aristotle's intuitions.

Some commentators (e.g. Lukasiewicz, 1967; Patzig, 1968; Hintikka, 1973) concluded that the combination of these statements just doesn't make any sense and that Aristotle must have been confused. Others, however, tried to account for these claims by looking for a consistent semantics of Aristotle's system (e.g. Thomason, 1993; Uckelman & Johnston, 2010). The most interesting of these latter accounts build on the idea that Aristotle's modal syllogistics was based on his metaphysics and philosophy of science (e.g. Rescher [1964], Van Rijen [2012], Patterson [2002], Malink [2013], Vecchio [2016])[3] Unfortunately, most of these authors have difficulty making many predictions of valid modal syllogistic reasoning that correspond with Artstotle's intuitions. Recently, however,

[3]Some (Van Rijen [2012]) have claimed that $Ba^{\Box}C$ can hold only if '$B$' is a substance term. This won't quite be enough (cf. Rini et al. [1998], Malink [2013]. Malink [2013] demands on top that a substance term can only be predicated of another substance term. We take this to follow naturally from a causal view.

Malink [2013] has shown that it is actually possible to come up with a systematic analysis of modal syllogistic sentences such that it gives rise to predictions almost exactly in accordance with Aristotle's claims.[4] His analysis, however, is quite involved. One wonders whether a simpler analysis is not possible. As mentioned earlier, we think such a simpler analysis is possible, if we make use of a *causal* analysis of modal categorical statements.

In this paper we will argue for a *causal* analysis of Aristotle's modal claims. We will argue that this fits well with Aristotle's analysis of demonstrative inferences in the Posterior Analytics, and that in this way we can account in a relatively straightforward way for several puzzling aspects of Aritotle's system of modal syllogisms presented in his Prior Analytics.

## 5.3 Causal analysis and Aristotelian demonstrations

### 5.3.1 Causal dependence and causal models

Consider the following two sentences:

(15) a. Aspirin causes headaches to diminish.

b. Aspirin relieves headaches.

Intuitively, (15a) says that there exists a causal connexion between Aspirin and diminishing headaches: the intake of Aspirin *tends to* diminish headaches. Remarkably, (15a) seems to express the same content as the *generic* sentence (15b). This strongly suggests that also the generic sentence (15b) should be given a causal analysis. Thus, not only (15a), but also (15b) expresses the fact that particular intakes of Aspirin *tend* to cause particular states of headache to go away, because of *what it is* to be Aspirin. Or, as we will say, because of the *causal power* of Aspirin to relieve headaches.

Causality is a kind of dependence. A number of authors have recently argued for a dependency analysis of conditionals, which is most straightforwardly done using probabilities: $C$ depends on $A$ iff $P(C|A) > P(C)$.[5] However, Douven (2008) has argued that dependence is not enough, 'If $A$, then $C$' is acceptable only if both $P(C|A) > P(C)$ *and* $P(C|A)$ are high.

We can implement Douven's proposal by requiring that $P(C|A) - P(C|\neg A)$ is close to $1 - P(C|\neg A)$. Since $P(C|A) > P(C)$ iff $P(C|A) > P(C|\neg A)$, we can demand that the conditional is acceptable iff $\frac{P(C|A)-P(C|\neg A)}{1-P(C|\neg A)}$ is high. This can only be the case if both $P(C|A) - P(C|\neg A)$ and $P(C|A)$ are high, so it derives Douven's demands.

[4] Vecchio [2016], building on Malink [2013], even slightly improves on Malink's predictions.
[5] For a discussion of some qualitative variants, see Spohn (2013) and Rott (ms).

The measure $\frac{P(C|A)-P(C|\neg A)}{1-P(C|\neg A)}$ is interesting from a causal perspective. Especially among philosophers dissatisfied with a Humean metaphysics, **causal powers** have recently become en vogue (again). Indeed, a growing number of philosophers (Harré & Madden, 1975; Cartwright, 1989; Shoemaker 1980; Bird 2007) have argued that causal powers, capacities or dispositions are the truth-makers of laws and other non-accidental generalities. Cheng (1997) hypothesises the existence of stable, but unobservable causal powers (Pearl [2000] calls them 'causal *mechanisms*') $p_{ac}$ of (objects or events of kind) $A$ to produce $C$. Cheng then *derives* a way how this objective but unobservable power can be estimated by an observable quantity, making use of standard probability theory and assuming certain natural independence conditions. It turns out that this quantity is exactly the above measure: $p_{ac} = \frac{P(C|A)-P(C|\neg A)}{1-P(C|\neg A)}$. Cheng's notion has been used for the analysis of conditionals, generics and disposition statements, in Schulz and Rooij [2019], van Rooij and Schulz [2019].

Dispositions and causal powers are things that (kinds of) objects have, independently of whether they show them. It is standardly assumed, though, that these (kinds of) objects *would* show them, if they *were* triggered sufficiently. Thus, there should be a relation with counterfactuals. Pearl (2000) provides a causal analysis of counterfactuals. He shows, however, that his (intervention-based) notion of 'probability of causal sufficiency of $A$ to produce $C$', abbreviated by $PS^C_A$, can be estimated under natural conditions by the same observable quantity $\frac{P(C|A)-P(C|\neg A)}{1-P(C|\neg A)}$ as Cheng's notion of causal power.

To see how he derives this quantity, we need to introduce causal models. We will use causal models to represent causal and counterfactual relationships. With Pearl [2000] we assume that such models represent a collection of 'mechanisms', a set of stable and autonomous relationships, represented by equations. A causal model, $\mathcal{M}$, is a triple, $\langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$, where $\mathcal{U}$ and $\mathcal{V}$ are disjoint finite sets of exogenous and endogenous variables. Let the members of $\mathcal{U}$ and $\mathcal{V}$ can be listed as $U_1, ..., U_n$ (written as $\overrightarrow{U}$ for short) and $V_1, ..., V_m$ (written as $\overrightarrow{V}$ for short) respectively. $\mathcal{F}$ is a set of mappings: each $f_i \in \mathcal{F}$ is a mapping which gives the value of $V_i$ given the values of all other variables in $\mathcal{U} \cup \mathcal{V}$. While the values of the *exogenous* variables $\mathcal{U}$ are determined by factors outside the model, the values of the *endogenous* variables $\mathcal{V}$ are determined by the values of variables in the model, i.e., by $\mathcal{U} \cup \mathcal{V}$. What $\mathcal{U}$ represents depends on the application of the causal model. It can represent a set of situations, but also, for instance, a set of objects. Depending on the application, an assignment to $\mathcal{U}$, namely $\overrightarrow{U} = \overrightarrow{u}$, can thus represent, for instance, a particular situation, or object (or vector of situations or objects). Such an assignment of $\mathcal{U}$ uniquely determines the values of all the endogenous variables. Thus, if $X$ is an endogenous variable, $X(\overrightarrow{u})$ gives the value of $X$ if $\overrightarrow{U} = \overrightarrow{u}$. The set of mappings $\mathcal{F}$, finally, represent the mechanisms, or causal dependencies. More particularly, each function $f_i$ can be written as an equation

$$(16)\ v_i \quad = \quad f_i(pa_i, u_i)$$

where $pa_i$ denotes the values of the endogenous variables that are the *parents* of $V_i$, and where the $u_i$ are the set of exogenous variables on which $f_i$ depends.

For the representation of hypothetical changes, or interventions, Pearl (2000) makes use of *submodels*. A submodel $\mathcal{M}_x$ of $\mathcal{M}$ is the causal model $\langle \mathcal{U}, \mathcal{V}, \mathcal{F}_x \rangle$, where $x$ is a particular value of endogenous variable(s) $X$, and where $F_x$ is just like $F$, except that all functions $f_x$ that correspond to members of $X$ are replaced by constant function $X = x$. Intuitively, $\mathcal{M}_x$ represents the minimal change from $\mathcal{M}$ required to make $X = x$ true for any $u \in U$. If $X$ and $Y$ are variables in $V$, the counterfactual 'The value that $Y$ would have obtained, had $X$ been $x$' is interpreted as denoting $Y_x(u)$. Intuitively, this will just be $f_y(x, u)$, in case $X$ is the only parent of $Y$.

In this paper we will make use of a *probabilistic causal model*. This is a pair $\langle \mathcal{M}, P(u) \rangle$, where $\mathcal{M}$ is a causal model, and $P(u)$ is a probability function defined over the domain $U$. Because each endogenous variable is a function of $U$, $P(u)$ completely defines the probability distribution over the endogenous variables as well. If $Y$ is a variable in $V$, we will abbreviate $P(Y = y)$ from now on by $P(y)$. The latter is determined as follows:

$$(17)\ P(y) \quad := \quad \sum_u P(u) \times \begin{cases} 1, & \text{if } Y(u) = y \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, if $X$ is also a variable in $V$, we will abbreviate $P(Y_x = y)$ by $P(y_x)$. The latter is determined as follows:[6]

$$(18)\ P(y_x) \quad := \quad \sum_u P(u) \times \begin{cases} 1, & \text{if } Y_x(u) = y \\ 0, & \text{otherwise.} \end{cases}$$

One of the most appealing features of calculating $P(y_x)$ as proposed by Pearl is that in this way we can also determine the probability $Y = y$ would have after an invention that would make $x$ true, if $x$ and $y$ are, in fact, not true. Thus, on Pearl's analysis one can easily determine $P(y_x \mid \neg x, \neg y)$. This latter notion is calculated as follows:[7]

$$(19)\ P(y_x \mid \neg x, \neg y) := \frac{P(y_x, \neg x, \neg y)}{P(\neg x, \neg y)} = \sum_u P(u \mid \neg x, \neg y) \times \begin{cases} 1, & \text{if } Y_x(u) = y \\ 0, & \text{otherwise.} \end{cases}$$

[6] For those who are not familiar with causal models, it might help to think of $P(y_x)$ as the probability of $y$ after *imaging* $P$ by $x$, as proposed by Lewis (1976), if $X$ and $Y$ correspond to variables associated with propositions.

[7] For convenience, we will sometimes use a comma, instead of logical '$\wedge$' below.

The probability of causal sufficiency of $A$ to produce $C$, abbreviated by $PS_A^C$, is thus determined as $P(C_A \mid \neg C, \neg A) = \frac{P(C_A, \neg C, \neg A)}{P(\neg C, \neg A)}$.

We will now show (following Pearl, 2000, chapter 9) that under natural conditions this reduces to $\frac{P(C|A)-P(C|\neg A)}{1-P(C|\neg A)}$, i.e., to Cheng's notion of causal power. To do so, we will first assume (i), a *consistency assumption* used for counterfactuals,

(i) $A \Rightarrow (C_A = C)$.

This assumption is natural: if $A$ already holds, an intervention to make $A$ true leaves everything as is.[8] Pearl also assumes a notion of *exogeneity*, i.e., that $C_A$ is *independent* of learning $A$ (and thus also that $\neg C_{\neg A}$ is *independent* of $\neg A$).

(ii) A variable $A$ is said to be **exogenous** relative $C$ in model $M$ iff $P(C_A \wedge C_{\neg A}|A) = P(C_A \wedge C_{\neg A})$.

Pearl's assumption that $A$ is exogenous to $C$ is very similar to Cheng's (1997) assumption that the potential causes of $C$ are *independent* of one another (the Noisy-OR assumption). It rules out that learning $A$ influences the probability of $C$ via an indirect way, for instance that if $B$ is another potential cause of $C$, there is a common cause of $A$ and $B$.

Making use of these two assumptions, we can make the following derivation:

$$
\begin{aligned}
\frac{P(C_A, \neg A, \neg C)}{P(\neg A, \neg C)} &= \frac{P(C_A, \neg A, \neg C_{\neg A})}{P(\neg A, \neg C)} && \text{by (i)} \\
&= \frac{P(C_A, \neg C_{\neg A}) \times P(\neg A)}{P(\neg A, \neg C)} && \text{by (ii)} \\
&= \frac{P(C_A \wedge \neg C_{\neg A}) \times P(\neg A)}{P(\neg C \wedge \neg A)} \\
&= \frac{P(C_A \wedge \neg C_{\neg A})}{P(\neg C / \neg A)} && \text{because } P(B \wedge A) = P(B/A) \times P(A) \\
PS_A^C &= \frac{P(C_A \wedge \neg C_{\neg A})}{1 - P(C/\neg A)}
\end{aligned}
$$

Next, we will derive that $P(C_A \wedge \neg C_{\neg A}) = P(C_A) - P(C_{\neg A})$, on the additional assumption of *monotonicity*:

(iii) $C$ is **monotonic** relative to $A$ iff for all $u$: $C_A(u) \geq C_{\neg A}(u)$.

Notice that because of monotonicity, if $C_{\neg A}$ is true, then $C_A$ is also true, and thus $\neg C_A$ is false. Thus, $C_{\neg A} \wedge \neg C_A$ cannot be true.

[8] If we would analyse the counterfactual $A \mathrel{\square\!\!\rightarrow} C$ by $C_A$, this consistency rule would validate the inference $A, C \therefore A \mathrel{\square\!\!\rightarrow} C$. This inference rule is accepted by almost everyone working on counterfactuals, although, to be honest, not by everyone.

$$\begin{aligned} C_A &= C_A \wedge (C_{\neg A} \vee \neg C_{\neg A}) \qquad \text{because } C_{\neg A} \vee \neg C_{\neg A} = \top \\ &= (C_A \wedge C_{\neg A}) \vee (C_A \wedge \neg C_{\neg A}) \\ C_{\neg A} &= C_{\neg A} \wedge (C_A \vee \neg C_A) \\ &= (C_{\neg A} \wedge C_A) \vee (C_{\neg A} \wedge \neg C_A) \\ &= C_{\neg A} \wedge C_A \qquad \text{because } C_{\neg A} \wedge \neg C_A \text{ is false (monotonicity)} \end{aligned}$$

By substituting $C_{\neg A}$ for $C_{\neg A} \wedge C_A$ in the elaboration of $C_A$, we get $C_A = C_{\neg A} \vee (C_A \wedge \neg C_{\neg A})$. Because $C_{\neg A}$ is incompatible with $\neg C_{\neg A}$, and thus with $C_A \wedge \neg C_{\neg A}$, it follows that $P(C_A) = P(C_{\neg A}) + P(C_A \wedge \neg C_{\neg A})$, and thus $P(C_A \wedge \neg C_{\neg A}) = P(C_A) - P(C_{\neg A})$. Thus, following Pearl we have derived that

$$(20)\ PS^C_A = \frac{P(C_A)-P(C_{\neg A})}{1-P(C/\neg A)}.$$

By using the rule $A \Rightarrow (C_A = C)$ (and thus also $\neg A \Rightarrow (C_{\neg A} = C)$), we can derive with exogeneity that $P(C_A) = P(C_A|A) = P(C|A)$ and $P(C_{\neg A}) = P(C_{\neg A}|\neg A) = P(C|\neg A)$. But this means that under the natural conditions of (i) consistency, (ii) exogeneity and (iii) monotonicity, we have established that

$$(21)\ PS^C_A = \frac{P(C|A)-P(C|\neg A)}{1-P(C/\neg A)}.$$

Thus, $PS^C_A$ can be thought of as the causal power of $A$ to produce $C$, i.e., $p_{ac}$. Notice that if all involved causal powers have value 1, a sequence of such causal powers is **transitively closed**: if $PS^B_A = 1$ and $PS^C_B = 1$, then also $PS^C_A = 1$. Obviously, also $PS^A_A = 1$, meaning that causal power is **reflexive**, and that demanding $PS$ to be 1 gives rise to a pre-order. Interesting about the probabilistic measure $\frac{P(C|A)-P(C|\neg A)}{1-P(C|\neg A)}$ is that it has its *maximal value* 1 just in case $P(C|A) = 1$ and $P(C|\neg A) \neq 1$.[9] Similarly, we predict that $PS^{\neg C}_A = 1$ and $p_{a\neg c} = 1$ holds only if $P(C|A) = 0$ and $P(C|\neg A) \neq 0$. Interestingly, $p_{a\neg c}$ corresponds with Cheng's (1997) notion of *causal power* of $A$ to *prevent* $C$. We propose that these notions might help us to provide a natural semantics for Aristotle's modal categorical sentences in order to illuminate Aristotle's hard to understand system of modal syllogisms.

### 5.3.2 A causal analysis of Aristotelian demonstrations

Many dialogues of Plato focus on questions of the form 'What is $X$?', where $X$ is typically some moral property like virtue or courage, a natural kind of

[9] Of course, the causal notions $PS^C_A$ and $p_{ac}$ demands this as well in case their values are 1, but in addition they demand that $A$ is a cause of $C$, and not that $A$ is uniquely caused by $C$. If we limit ourselves to values that are 1 or not, the probabilistic measure is antisymmetric, and thus gives rise to a partial order.

thing like human, or water, or a mathematical object like a triangle. A good answer to this kind of question must consist of a set of features all and only all individuals of type $X$ have. Aristotle, a pupil of Plato, was interested in the same kind of questions. But he also was more ambitious. If all (and only all) individuals or objects of type $X$ share certain features, Aristotle also wanted to know *why*. Indeed, for Aristotle, scientific inquiry is an attempt to answer 'why' questions. A scientific explanation of a fact about the world consists of a valid syllogistic argument with some fundamental true claims as its premises and this fact as the conclusion. But not any old valid syllogism would do, for the premises must express *fundamental* true claims. A valid syllogism that satisfies this extra requirement Aristotle calls a *demonstration*. A typical Aristotelian demonstration is the following:

(22) a. All humans are animals.

b. All animals are living things.

c. Therefore, all humans are living things.

In this demonstration, the two premisses are taken to express essential features of humans and animals, respectively. They follow from Aristotle's theory of *real definitions* of objects of type $X$ in terms of (i) an immediately higher type $Y$, and a differentia $Z$. If $X$ is 'human', for instance, then $Y$ would be 'animal', and $Z$ would be 'rational': a man is a rational animal. Thus, in 'All humans are animals', 'being animal' is essentially predicated of humans, and the first premise of the above syllogisms can be expressed by $Sa^{\Box}P$. However, not all true sentences of the form $Sa^{\Box}P$ can be read off directly from Aristotle's theory of real definitions. Some have to be indirectly derived. This is what happens in the above syllogism. In the above syllogistic argument, the premisses can be directly read off from Aristotle's theory of definition, but to reach the conclusion an additional argument is needed. This is provided by the syllogism, that can be stated as being of the form $Aa^{\Box}B, Ba^{\Box}C \therefore Aa^{\Box}C$. For Aristotle, this argument *explains why* humans are living things. The argument turns a fact into a *reasoned fact*.[10]

What has this all to do with causality? Well, Aristotle had a somewhat wider notion of causality than many moderns have. For him, it is necessary for humans to be able to learn grammar. But being able to learn grammar is not an essential property of humans or of any higher kind. It just *causally follows by necessity* from being rational (according to Aristotle). Thus, even though all and only all objects of type $X$ have feature $f$ and $g$, it can be that one of the features is still only a derived feature, causally derived.

[10]For much more detailed and sophisticated analyses of Aristotelian demonstration see Crager (2015) and Vecchio (2016).

So far, it seems that scientific demonstrations must consist of two premisses that are both necessary. But this is not exactly what Aristotle seems to assume. In fact, in his *Posterior Analytics* Aristotle discusses the following two valid syllogisms:

(23) a. All (the) planets are near the earth

b. All objects that are near the earth do not twinkle

c. Therefore, (all) the planets do not twinkle.

and

(24) a. All (the) planets do not twinkle.

b. All objects that do not twinkle are near the earth.

c. Therefore, (all) the planets are near the earth.

In these arguments, the premises (23a) and (24b) are not taken to express necessary truths. Although the second syllogism is not taken to be a scientific demonstration, Aristotle claims that the first syllogistic inference is. It leads to a 'reasoned fact', because the middle term 'being near the earth' *causally explains* the conclusion, something that is not the case for the middle term in the other inference 'objects that do not twinkle'. If we would translate the above arguments in modal syllogistic terms, they would be of the forms $AaB, Ba^{\Box}C \therefore Aa^{\Box}C$ and $Aa^{\Box}B, BaC \therefore Aa^{\Box}C$, respectively. Note that they are thus of types Barabara LXL and Barbara XLL, respectively.[11] Note also that in his Prior Analytics, Aristotle took only the first type of argument valid. So, there seems to be a close relation between what Aristotle claims in his two Analytics.

## 5.4 Causality and modal syllogisms

Causal links need not only connect propositions, they can connect properties, or features, as well. In fact, Danks [2014] argues that all prominent theories of concepts could be represented by graphical causal models. Although not explicitly discussed, the essentialists' version is one:: features of birds are connected (and thus caused) in various strengths to the essence of the kind, i.e., by what it is to be a bird.

Let us now come back to the question what the natural interpretation of Aristotle's *modal* syllogistics is.

Recall that $Ba^{\Box}C$ means that all $B$s are *necessary/essentially* $C$ and that Aristotle claimed that the following modal syllogisms are valid and invalid, respectively:

[11] According to Vlecchio (2016), the argument in (9) explains why planets do not twinkle, by using a fact is which part of the nominal definition of a planet ('being near the earth'), but which is not a part of its real definition.

| | | | |
|---|---|---|---|
| ***(i)*** | $Ba^{\Box}C, Aa^{\Box}B \therefore Aa^{\Box}C$ | Valid | Barbara LLL |
| ***(ii)*** | $Ba^{\Box}C, AaB \therefore Aa^{\Box}C$ | Valid | Barbara LXL |
| ***(iii)*** | $BaC, Aa^{\Box}B \therefore Aa^{\Box}C$ | Invalid | Barbara XLL |

Similarly, Aristotle claims that the following modal syllogism is valid, where $Be^{\Box}C$ means that by (*de re*) necessity no $B$ is a $C$:

| | | | |
|---|---|---|---|
| 4. | $Be^{\Box}C, AaB \therefore Ae^{\Box}C$ | Valid | Celarent LXL |

Moreover, Aristotle claims that not only conversion inference 5 is valid, but that the same holds for the modal conversion inferences 6 and 7:

| | | |
|---|---|---|
| 5. | $BeC \therefore CeB$ | Valid |
| 6. | $Be^{\Box}C \therefore Ce^{\Box}B$ | Valid |
| 7. | $Ba^{\Box}C \therefore Bi^{\Box}C$ | Valid |

We claim that Aristotle's claims make perfect sense once we understand $Ba^{\Box}C$ as causally explaining *why* $C$. More in particular, we would like to say that $Ba^{\Box}C$ just means that $B$ has complete causal power to make $C$ to hold, i.e., $PS_B^C = 1$ (or $p_{bc} = 1$) and that $Be^{\Box}C$ just means that $B$ and $C$ have complete causal powers to prevent each other to hold, i.e., $PS_B^{\neg C} = 1$ and $PS_C^{\neg B} = 1$ (or $p_{b\neg c} = 1$ and $p_{c\neg b} = 1$).[12] So we argue that Aristotle's modal syllogisms are based on causal principles. As noted above, for $PS_B^C = 1$ (or $p_{bc} = 1$) to hold, it must be that (i) $P(C|B) = 1$ and (ii) $P(C|\neg B) \neq 1$. We will assume the probabilistic analysis of non-modal categorical sentences as mentioned in section 2. Thus, $BaC$ will be true iff $P(C|B) = 1$.

Inference 1 is valid on this interpretation, because if the premisses are true the following will hold (i) $P(C|B) = 1$, (ii) $P(C|\neg B) \neq 1$, (iii) $P(B|A) = 1$ and (iv) $P(B|\neg A) \neq 1$. Obviously, by (i) and (iii) it follows that $P(C|A) = 1$. From (ii) and (iv) it follows that (a) there are some $\neg C$s among the $\neg B$s, and (b) that there are some $\neg B$s among the $\neg A$s. By (a) and (b) this means that $P(C|\neg A) \neq 1$. Thus, $P(C|A) = 1$ and $P(C|\neg A) \neq 1$ which means that $Aa^{\Box}C$.

Inference 2 is also valid on this interpretation, because if the premisses are true it means that the following will hold (i) $P(C|B) = 1$, (ii) $P(C|\neg B) \neq 1$ and (iii) $P(B|A) = 1$. Obviously, by (i) and (iii) it follows again that $P(C|A) = 1$. From (ii) it follows that there are some $\neg C$s among the $\neg B$s. But because $AaB$, it holds that all $\neg B$s are $\neg A$s, and thus there must also be some $\neg C$s among the $\neg A$s. Thus, $P(C|A) = 1$ and $P(C|\neg A) \neq 1$ which means that $Aa^{\Box}C$.

[12] Aristotle's (hyperintensional) distinction between necessity and essentiality suggests that the analysis of $Ba^{\Box}C$ as $p_{bc} = 1$ is still too coarse-grained. Notice, however, that even if $B$ and $C$ are necessary co-extensive, it will typically be (causally speaking) that either $p_{bc} = 1$ and $p_{cb} = 0$, or $p_{bc} = 0$ and $p_{cb} = 1$. We take the former to be the case if $B$ is a substantive term and $C$ an adjectival one. We assume, however, that $B$ and $C$ can *prevent* each other (to account for $Be^{\Box}C \therefore Ce^{\Box}B$).

Inference 3, however, is not valid. The important thing to observe is that this is just an instance of **right weakening**,[13] an inference which should (and does) **not** hold on our causal analysis. In particular, the inference has a counterexample, in case the domain consists only of $C$ individuals.

Similarly, we can account for Aristotle's intuition that inference 4 is valid. Using the above interpretation of non-modal statements, we account for inference 5. As for the validity of inference 6, notice that $Be^{\Box}C$ holds iff (i) $P(C|B) = 0$ and (ii) $P(C|\neg B) \neq 0$. From (i) it immediately follows that also (a) $P(B|C) = 0$, if $\Diamond C$. But that $\Diamond C$ follows immediately from (ii). But from (i) it also follows that (b) there is a $B$ that is $\neg C$ (for otherwise $P(C|Ba)$ would not be defined). From (a) and (b) it immediately follows that $Ce^{\Box}B$ holds. (Actually, we have to consider $P(B|\neg C) \neq 0$ as well, but that is analogue to the other case.)[14] As for inference 7, this immediately follows from the semantic analysis of statements like $Ci^{\Box}B$ to be given in a minute.

Our predictions agree with all Aristotle's claims of (in)validities of universal modal syllogisms with modality $\Box$. For instance, we correctly predict Aristotle's claimed validity of Cesare LXL, Camestres XLL, and his claim of invalidity of Camester LXL. The latter one – $Ba^{\Box}A, CeA \not\models Be^{\Box}C$ – is particularly interesting. It is easy to see that this inference would be predicted as valid, if we analysed $CeA$ as true iff $P(A|C) = 0$, which presupposes that $P(C) \neq 0$. However, we have analysed $CeA$ as true iff $P(C \cap A) = 0$, and on this interpretation Camestres LXL is *not* predicted to be valid, in accordance with Aristotle's intuitions. More in particular, our analysis makes the right predictions for the modal Barbara and Celarent syllogisms of the first figure.

As for the second figure, and limiting ourselves to universal statements, we have to explain why (according to Aristotle)

(25) a. $Ae^{\Box}B, CaB \models Ce^{\Box}A$ Cesare LXL

b. $AeB, Ca^{\Box}B \not\models Ce^{\Box}A$ Cesare XLL

and

(26) a. $Aa^{\Box}B, CeB \not\models Ce^{\Box}A$ Camestres LXL

b. $AaB, Ce^{\Box}B \models Ce^{\Box}A$ Camestres XLL

As for (25a), this follows immediately from our semantics, assuming that $Ae^{\Box}B$ holds if $p_{a\neg b} = 1$ and $p_{b\neg a} = 1$. For (25b) this follows because $BeA$ is true $P(B \cap C) = 0$. As for (26a). This doesn't follow, because it is not guaranteed that $P(C|\neg A) \neq 0$, which makes the conclusion false. Inference (26b) is immediately verified. There are no other modal syllogisms with only universal statements of

[13] In condidtional terms, right weakening means that if $A \Rightarrow B$ and $B \models C$, then also $A \Rightarrow C$.
[14] Perhaps it is better to interpret $Be^{\Box}C$ as $\neg(Bi^{\Box}C)$.

the second figure to be checked, and we don't know about Aristotle's intuitions on only 'universal' modal syllogisms of the fourth figure (Cameses$_4$), so these are all predicted in accordance with Aristotle's intuition.

As for modal syllogisms with non-universal sentences, Aristotle claims (27a) (of the first figure) to be valid, but (27b) not to be so:

(27) a. $Ba^{\Box}A, CiB \models Ci^{\Box}A$ Darii LXL

b. $BaA, Ci^{\Box}B \not\models Ci^{\Box}A$ Darii XLL

To account for this, we first need to know what makes $Ci^{\Box}B$ true. In counterfactual terms, we propose that this is true iff $\exists x : xaC, \exists D : xa^{\Box}D$ and $P(B_D|\neg B, \neg D) = 1$,[15] where $xaC$ is the singular categorical sentence that (all) $x$ is $C$, and $xa^{\Box}D$ the singular categorical sentence that (all) $x$ is necessary $D$. This is enough to make Darii LXL valid. However, this meaning rule makes Darii XLL invalid, as desired. Notice that in non-counterfactual terms, the symmetric version of our interpretation of $Ci^{\Box}B$ comes down to the following: $Ci^{\Box}B$ is true iff $\exists x \in C \cap B$ and $P(B|\neg x) \neq 1$ or $P(C|\neg x) \neq 1$. Similarly, $Co^{\Box}B$ is true iff $\exists x, y : xaC, yeC$ and $P(B|x) \neq 1$ and $P(B|y) = 1$.[16] This latter more specific interpretation rule for $Co^{\Box}B$ (using $y$ such that $yeC$ instead of $\neg x$) is required to account, for instance, for Aristotle's claimed invalidity of Barocco XLL.[17]

Aristotle also claims a distinction between the following syllogisms also of the first figure:

(28) a. $Be^{\Box}A, CiB \models Co^{\Box}A$ Ferio LXL

b. $BeA, Ci^{\Box}B \not\models Co^{\Box}A$ Ferio XLL

Inference (28a) follows immediately if we analyse $Co^{\Box}A$ as true iff $\exists x : xaC, \exists D : xa^{\Box}D$ and $P(\neg A_D|B, \neg D) = 1$. There is an easy counterexample to (28b), again due to the fact that the conclusion $Co^{\Box}A$ demands that there is at least one $A$, while premise $BeA$ can be true without there being such an $A$.

Aristotle didn't give his opinion on every possible syllogism which involves sentences with necessity modals. In fact, he limited himself to syllogisms that (i) have a necessity modal in the conclusion, (ii) are of the first three figures and (ii) that are valid without any modal. Still, there are 6 valid syllogisms in each figure, and 3 possible combinations where at least one of the premises has a necessity modal. Of those 54 syllogisms, Aristotle expressed his opinion on 42 of those modal syllogisms.[18] 23 of those syllogisms he counted valid, and the

[15] Or better, the following symmetric version to account for conversion $Ci^{\Box}B \models Bi^{\Box}C$, $Ci^{\Box}B$ is true iff $\exists x, \exists D : xa^{\Box}D$ and (i) $xaC$ and $P(B_D|\neg B, \neg D) = 1$ or (ii) $xaB$ and $P(C_D|\neg C, \neg D) = 1$.

[16] Formulated more sophisticated: $Ci^{\Box}B$ is true iff $\exists x : xaC$ and $xaB$ and ($p_{xb} = 1$ or $p_{xc} = 1$).

[17] To be clear, a much simpler interpretation rule for $Co^{\Box}B$ is possible, if $AaB$ would be interpreted as true iff $A \subseteq B$. Then we could just say that $Co^{\Box}B$ is true iff $\exists X : xaC$ and $p_{x\neg b} = 1$.

[18] We base ourselves here completely on appendix A of Malink [2013]

others non-valid. He looked at 14 syllogisms where all categorical sentences involved had a necessity modal, such as Barbara LLL, and he counted all of them as valid. We can check that all such modal syllogisms are valid on our analysis as well. Let us go to one of the more challenging ones to explain: Darii LLL, $Ba^{\Box}A, Ci^{\Box}B \models Ci^{\Box}A$. The first premise means that $PS^A_B = 1$. According to the second premise, $\exists x : xaC, \exists D : xa^{\Box}D$ and $PS^B_D = 1$. Because if $PS^A_B = 1$ and $PS^B_D = 1$, it follows by transitivity that also $PS^A_D = 1$. It follows that thus $\exists x : xaC, \exists D : xa^{\Box}D$ and $PS^A_D = 1$, which means that conclusion $Ci^{\Box}A$ is true.

As for the other 30 modal syllogisms of this type that Aristotle considered, we checked them as well, and our analysis predicts in accordance with Aristotle's intuitions. Thus, our analysis makes predictions exactly in accordance with Aristotle's claims of (in)validity for **every** modal syllogism Aristotle's explicitly discussed! We think this is quite remarkable.

We haven't checked our predictions for *all* 16.384 modal syllogisms, though. In fact, we didn't check any syllogism with the possibility and contingency modals that Aristotle also discussed. In this paper we did not even propose meanings of such sentences. Of course, for the standard possibility modal, a natural analysis suggests itself:

$$\text{(29)} \qquad Aa^{\Diamond}B \equiv \neg(Ao^{\Box}B) \qquad Ae^{\Diamond}B \equiv \neg(Ai^{\Box}B)$$

$$Ai^{\Diamond}B \equiv \neg(Ae^{\Box}B) \qquad Ao^{\Diamond}B \equiv \neg(Aa^{\Box}B)$$

But it remains to be seen whether this analysis gives rise to predictions that accords with Aristotle's intuitions. It is even less clear whether we can account for Aristotle's claims involving the contingency modal, $\Delta$, a task that is perhaps the most challenging. Striker (1985) argues, though, that sentences like $Aa^{\Delta}B$ should be interpreted basically as generic sentences, where $B$ applies *by nature*, or *for the most part*, to $A$. Interestingly, this suggestion would be quite in line with van Rooij & Schulz (2020), according to which generic sentences of the form '$A$s are $B$' are interpreted as having high causal power, i.e. $p_{ab} \approx 1$. But it is more natural to interpret $Aa^{\Delta}B$ as $\forall x \in A : \neg\exists D : xa^{\Box}D$ and ($Da^{\Box}B$ or $De^{\Box}B$) and $Ai^{\Delta}B$ as $\neg\exists x \in A : \exists D : xa^{\Box}D$ and ($Da^{\Box}B$ or $De^{\Box}B$) to account for Aristotle's claims that $Aa^{\Delta}B$ is equivalent with $Ae^{\Delta}B$ and $Ai^{\Delta}B$ with $Ao^{\Delta}B$, and that not only $Ai^{\Delta}B$ is equivalent with $Bi^{\Delta}A$, but also that $Ao^{\Delta}B$ is equivalent with $Bo^{\Delta}A$. We don't know whether with this interpretation we can account for all of Aristotle's intuitions w.r.t. modal syllogisms involving $\Delta$. But the smoothness of our explanation of Aritotle's above intuitions makes one optimistic.

But there is further ground for optimism. Malink [2013] and Vecchio [2016] have recently shown how to account for most (if not all) of the Aristotle's claims about modal syllogisms making use of *essences*. $Ba^{\Box}C$ is true iff all $B$s are $C$ *in virtue of* what it is to be a $B$. But that is exactly how we think of our own proposal as well.

## 5.5 A challenge: counterexamples to Barbara LXL?

We have shown in the previous section that our causal power analysis can account for why the modal syllogism Barbara LXL, $Ba^{\Box}C, AaB \therefore Aa^{\Box}C$ is valid, although Barbara XLL, $BaC, Aa^{\Box}B \therefore Aa^{\Box}C$, is not. We have seen that this can be shown if we analyse statements like $Ba^{\Box}C = 1$ by $\frac{P(C|B)-P(C|\neg B)}{1-P(C|\neg B)} = 1$ and $AaB = 1$ by $P(B|A) = 1$. We have also seen that the causal notions of causal power and $PS^C_A$ come down to this probabilistic notion under certain circumstances.

Although Aristotle claimed that Barbara LXL is valid, very soon (putative) counterexamples to this modal syllogisms were offered:[19]

(30) a. All litererats necessarily have knowledge, all men are litarate, thus all men necessarily have knowledge.

b. $Ba^{\Box}C, AaB \therefore Aa^{\Box}C$ Barbara LXL

In fact, Aristotle himself provided a (putative) counterexample to Celarent LXL himself.

(31) a. All ill people are necessarily not healthy, all men are ill, thus all men are necessarily not healthy.

b. $Be^{\Box}C, AaB \therefore Ae^{\Box}C$ Celarent LXL

Malink [2013] and Crager [2015] argue that these counterexamples can be explained away if we take seriously Aristotle's analysis of 'genuine predication' from Aristotle's *Categories*. The idea is that terms can denote sets of different *ontological types*: some denote *substances*, while others denote *qualities*. Just as each substance has an essence, this is also the case for each quality. However, denotations of the same type can only stand in a limited number of extensional relations with each other. For instance, for any two substances $A$ and $B$, it cannot be that $A \cap B \neq \emptyset$ without either $A \subset B$ or $B \subset A$. Beyond this *extensional* constraint, there lays a more important *intensional* constraint: if $A$ and $B$ are of the same ontological type, then, if $A \subset B$, then $Aa^{\Box}B$. Malink [2013] and Crager (2015) argue that Aristotle took Barbara LXL and Celarent LXL to be valid because he demanded that in a demonstration with a necessary conclusion, also the seemingly nonmodal premise (in our cases, the minor premise $AaB$) should be a case of genuine predication.

If Malink [2013] and Crager [2015] are correct, it means that valid modal syllogisms with a necessity modal in the conclusion should, in the end, all be of the form LLL. It also suggests that our explanation in the previous section of the validity of Barbara LXL and Celarent LXL will not be correct, for otherwise the

[19] For modern discussion, see van Rijen (1989),Rini et al. [1998], Van Rijen [2012], Malink [2013], Crager [2015].

(putative) counterexamples above would likely be genuine counterexamples. If we want to stick to our causal analysis, this suggests that instead of looking at the *extensional* notion $\frac{P(C|B)-P(C|\neg B)}{1-P(C|\neg B)} = 1$ for the analysis of $Ba^{\Box}C$ we should look at the *intensional* counterpart, $\frac{P(C_B)-P(C_{\neg B})}{1-P(C|\neg B)} = 1$, where intervention still plays an important role, and the counterfactual probabiltiy $P(B_A)$ is not reduced to the conditional probability $P(B|A)$. Indeed, on such an intensional analysis Barbara LXL, $Ba^{\Box}C, AaB \therefore Aa^{\Box}C$, would not be valid, because from $\frac{P(C_B)-P(C_{\neg B})}{1-P(C|\neg B)} = 1$ and $P(B|A) = 1$, we cannot conclude that $\frac{P(C_A)-P(C_{\neg A})}{1-P(C|\neg A)} = 1$.

We don't know, though, whether Malink's (2013) and Crager's (2015) interpretation of Aristotle is correct. For one thing, Malink [2013] himself already notes that Aristotle explicitly discusses modal syllogisms that he takes to be valid even though the nonmodal premise does not seem to involve genuine predication. But, of course, if Malink and Crager are not correct, we would have to explain away the above 'putative' counterexamples in another way. In fact, Vecchio (2016, chapter 1) argues that Aristotle himself explained away the (putative) counterexamples to Barbara LXL and the like in a more straightforward way than was suggested by Malink (2013): by demanding that the terms are interpreted in an omnitemporal way, which makes the non-modal premise false. Vecchio (2016, chapter 3) also argues explicitly that Aristotle used syllogisms of the form Barbara LXL in his analysis of scientific demonstrations in the *Posterior Analytics*, just as we suggested in section 3.2. Vecchio argues that Barbara LXL can be used to turn a *nominal* definition, 'Thunder is a noise in the clouds' (of form *AaB*) to a *real* definition 'Thunder is (necessarily) the extinguishing of fire in the clouds' (of form $Aa^{\Box}C$) via the essential major premise 'A noise in the clouds is (by necessity) the extinguishing of fire in the clouds' (of form $Ba^{\Box}C$).[20] Note that if Vecchio is right, our 'extensional' causal analysis might be on the right track after all.

[20]There exists an interesting analogue between this and the way natural kind terms receive their content according to the causal theory of reference: first a set of superficial properties is used to identify a set of things, and later having these superficial properties is explained by some essential properties all the things in the set have in common.

# Chapter 6

# Conclusions

## 6.1 Taking stock

The goal of the intellectual journey we took in this dissertation was to advance our understanding of causal relations, especially the way we talk and reason about them. Let us take stock and review how far we got with answering the concrete research questions formulated in the introduction. For this we will build on the conclusion sections of the four papers that form the core of the thesis: Barbero et al. [N.D] , Schulz and Xie [N.D], Schulz et al. [2019] and van Rooij and Xie [N.D].

The first question we set out to answer was "Can we build a qualitative formal system that can handle both causal and epistemic reasoning together?" In chapter 2 we made a first step towards integrating causal and epistemic reasoning. We developed an extension of causal models that is able to represent the epistemic state of an agent and talk about this epistemic state in its object language. The goal was to allow reasoning about causal knowledge. The model we presented in this chapter can express knowledge and knowledge update via a dynamic operator. In addition, we provide an axiomatization for the formal framework and prove that the system of inference is sound and complete. So, even though many questions are still open when it comes to integrating causal and epistemic reasoning, we made a first step in the right direction and laid down a sound foundation for future work.

Chapter 3 focused on the meaning of conditional sentences, in particular counterfactual conditionals. The goal of this chapter was to provide an account for the meaning of a particular group of counterfactual conditionals that turns out to be problematic for the standard interventionist approach to counterfactuals. These are counterfactuals that seem to allow for an epistemic reading. Just as in Chapter 2 we approached this problem by combining causal models with a representation of the epistemic state of the speaker. But in contrast to the framework in the previous chapter, we additionally added a representation of

belief. Based on this new notion of model we proposed a semantics for counterfactuals that still operates along the lines of the interventionist approach, but additionally takes into account how an intervention affects the beliefs of an agent. We show that this approach allows us to keep all the nice predictions of the interventionist approach, but can also account for epistemic counterfactuals. And in contrast to other solutions proposed in the literature our approach does not have to postulate a second, epistemic reading for these sentences.

In chapter 4 we focused on a different limitation of the interventionist approach to counterfactuals: they appear to predict wrong truth conditions for right-nested counterfactuals. According to a recent diagnosis of this problem made by Fisher the problem is caused by a particular feature of the interventionist approach, which he calls *strict interventionism*. To solve this problem we define a new notion of intervention that does not have the property of strict interventionism. This new notion of intervention does not change structural dependencies encoded in the causal model (i.e. the *laws*), but only affects the valuation of the relevant variables (the *facts* of the world or point of evaluation). We show that this way we can account for the problematic observations. In other words, also this question formulated in the introduction received a first answer.

The goal of the fifth chapter was to study the question of whether a causal analysis of conditionals can shed some light on Aristotle's modal syllogisms. In contrast to the previous two chapters, the focus moved now from counterfactual conditionals to indicative conditionals. By using the notion of causal power introduced by [Cheng, 1997], we defined a semantics for indicative conditionals that we then applied to their use in modal syllogisms. It turns out that the predictions our approach makes for modal syllogisms fit well with Aristotle's judgements concerning their validity. So, it seems that also here a causal approach to the meaning of conditionals – and more specifically the approach that we chose here – is very fruitful, now in advancing our understanding of Aristotle's writings.

Summing up, based on these results we can say that indeed combining work on causality, conditionals and epistemology proves very useful to address open issues at the intersection of these research areas. But, of course, the investigations reported on in this thesis can only be the beginning. As indicated in the different chapters there are various interesting questions and further connections that should be explored in future work. We will highlight some of them in the next section, building again on the discussion provided in the papers Barbero et al. [N.D], Schulz and Xie [N.D], Schulz et al. [2019] and van Rooij and Xie [N.D].

## 6.2 Questions for future work

There are still many open questions for future work. For instance, the formal frameworks that we developed in the chapters 2 and 3 do not yet allow us to study the interaction of multiple agents. The reason is that the symbolic language that was introduced is not able to express the epistemic states of different agents. But, of course, there exist languages in dynamic epistemic logic that have been designed to model multi-agent scenarios, allowing for higher-order reasoning and the representation of group attitudes. Building on this work it would be very natural to extend our language and models in Chapter 2 and Chapter 3 to a multi-agent setting. This should allow the formal system to represent the interaction of epistemic attitudes in groups of agents with respect to causal reasoning.

There are also some aspects of the way we formalised causal reasoning and causal knowledge that are not completely convincing. For instance, a central property of the notion of causal knowledge that we implemented is the following validity: if after making $X = x$ true, the agent knows $Y = y$, then an agent knows that forcing $X = x$ will lead to $Y = y$ (i.e. $K([X = x]Y = y)$). However, several natural examples indicate that this validity does not hold in general. The reason why the notion of causal knowledge we have implemented gives rise to this validity is that in our system an agent cannot gain any new information as a consequence of an intervention. This is certainly a premise that should be given up in the future. One solution that we are currently working on is to specify a set of "observables", which represent variables that an agent is able to observe during a change that happens in the external world. Thus observables allow agents to learn about interventions.

There were also some issues with the conceptual choices we made in Chapter 3. The epistemic causal models defined in this chapter represent not only knowledge but also belief. When we applied this notion of model to concrete examples we categorized all the information the agent has due to direct observation as knowledge and modelled expectations the agent has because of causal laws that she endorses as beliefs. But there could be other sources of information contributing to the knowledge or beliefs of an agent. What, for instance, about analytical laws? Shouldn't they affect what an agent knows? Furthermore, the causal laws and observations usually affect beliefs in complex ways. For example, if two facts are causally independent from each other and there are no facts observed by the agent, then the two facts should be epistemically independent as well (and additional observations about one of the facts should not change the belief concerning the other). This type of interaction between observation, causality and belief should be reflected in our formal account. We leave this as well for future work.

Our proposal in Chapter 4 made an important step forward in the direction of understanding right nested counterfactuals. We showed that by switching

from interventions on the causal dependencies to interventions on the valuation of the involved variables observations that have so far been unaccounted for can be explained. However, this does not mean that our account makes correct predictions for all right-nested counterfactuals. As pointed out in Chapter 4, some experimental results can neither be explained by the standard interventionist approach nor by our semantics. Also accounting for these more complex examples is left for future work.

And to mention an open end left in chapter 5, we have only checked the predictions of the proposed causal analysis for modal syllogisms without possibility and contingency operators. The semantics for sentences with a contingency modality still needs to be spelled out. Also this is a direction we would like to explore in future research.

While the interaction between epistemology, causality and conditionals may well have been a bit oversimplified in our account, the analysis in the chapters of this thesis does show that it is the right direction forward to solve many open problems.

# Bibliography


A. Baltag. To know is to know the value of a variable. *Advances in modal logic*, 11:135–155, 2016.

A. Baltag and B. Renne. Dynamic epistemic logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

A. Baltag and S. Smets. Probabilistic dynamic belief revision. *Synthese*, 165(2): 179–202, 2008a.

A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. *Logic and the foundations of game and decision theory (LOFT 7)*, 3:9–58, 2008b.

A. Baltag and J. van Benthem. A simple logic of functional dependence. *Journal of Philosophical Logic*, 2020.

A. Baltag, L. S. Moss, and S. Solecki. The logic of public announcements, common knowledge, and private suspicions. In I. Gilboa, editor, *TARK*, pages 43–56, San Francisco, CA, USA, 1998. Morgan Kaufmann. ISBN 1-55860-563-0. URL http://dl.acm.org/citation.cfm?id=645876.671885.

A. Baltag, I. Canavotto, and S. Smets. Causal agency and responsibility: A refinement of stit logic. In A. Giordani and J. Malinowski, editors, *Logic in High Definition. Trends in Logical Semantics*, pages ??–?? forthcoming.

F. Barbero and G. Sandu. Interventionist counterfactuals on causal teams. In B. Finkbeiner and S. Kleinberg, editors, Proceedings 3rd Workshop on *formal reasoning about Causation, Responsibility, and Explanations in Science and Technology,* Thessaloniki, Greece, 21st April 2018, volume 286 of *Electronic Proceedings in Theoretical Computer Science*, pages 16–30. Open Publishing Association, 2019. doi: 10.4204/EPTCS.286.2.

F. Barbero and G. Sandu. Team semantics for interventionist counterfactuals: observations vs. interventions. To appear in the Journal of Philosophical Logic, 2020.

F. Barbero and F. Yang. Counterfactuals and dependencies on causal teams: expressive power and deduction systems. accepted by Advances in Modal Logic 2020, forthcoming.

F. Barbero, K. Schulz, S. Smets, F. R. Velázquez-Quesada, and K. Xie. Thinking about causation: a causal language with epistemic operators. N.D.

B. Bergstein. What AI still can't do, 2020. https://www.technologyreview.com/2020/02/19/868178/what-ai-still-cant-do/ [Accessed: June 2020].

P. Blackburn, M. de Rijke, and Y. Venema. *Modal logic*. Number 53 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, UK, 2001. ISBN 0-521-80200-8. doi: 10.1017/CBO9781107050884.

R. Briggs. Interventionist counterfactuals. *Philosophical studies*, 160(1):139–166, 2012.

P. W. Cheng. From covariation to causation: A causal power theory. *Psychological review*, 104(2):367, 1997.

H. Chockler and J. Y. Halpern. Responsibility and blame: A structural-model approach. *J. Artif. Int. Res.*, 22(1):93–115, Oct. 2004. ISSN 1076-9757.

I. Ciardelli, L. Zhang, and L. Champollion. Two switches in the theory of counterfactuals. A study of truth conditionality and minimal change. *Linguistics and Philosophy*, 2018.

A. D. Crager. Meta-logic in aristotle's epistemology. *PhD diss., Princeton University*, 2015.

D. Danks. *Unifying the mind: Cognitive representations as graphical models*. Mit Press, 2014.

M. Dehghani, R. Iliev, and S. Kaufmann. Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, 27:55–85, 2012.

D. Edgington. Causation first: Why causation is prior to counterfactuals. In C. Hoerl, T. McCormack, and S. R. Beck, editors, *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology*, pages 230–241. Oxford University Press, 2011.

D. J. Etlin. *Desire, belief, and conditional belief*. PhD thesis, Massachusetts Institute of Technology, 2008.

R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about knowledge*. The MIT Press, Cambridge, Mass., 1995. ISBN 0-262-06162-7.

K. Fine. Review of Lewis (1973). *Mind*, 84:151–158, 1975.

T. Fisher. Causal counterfactuals are not interventionist counterfactuals. *Synthese*, 194(12):4935–4957, 2017.

A. Frank. *Context dependence in modal constructions*. PhD thesis, University of Stuttgart, 1997.

D. Galles and J. Pearl. An axiomatic characterisation of causal counterfactuals. *Foundations of Science*, 1:151–182, 1998a.

D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1):151–182, 1998b.

T. Gerstenberg, C. Bechlivanidis, and D. A. Legnado. Back on track: Backtracking in counterfactual reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 35(35)*. Cognitive Science Society, 2013.

C. Glymour and P. Spirtes. Latent variables, causal models and overidentifying constraints. *journal of Econometrics*, 39(1-2):175–198, 1988.

H. P. Grice. Logic and conversation. In *Syntac and Semantics 3: Speech acts*, pages 41–58. Elsevier, 1975.

J. Y. Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.

J. Y. Halpern. From causal models to counterfactual structures. *The Review of Symbolic Logic*, 6(2):305–322, 2013.

J. Y. Halpern. *Actual causality*. MIT Press, 2016.

S. Hansson. New operators for theory change. *Theoria*, 55(2):114–132, 1989.

E. Hiddleston. A causal theory of counterfactuals. *Noûs*, 39(4):. 632–657, 2005.

J. Hintikka. *Logic, language-games and information: Kantian themes in the philosophy of logic*. Clarendon Press Oxford, 1973.

C. Hitchcock. A tale of two effects. *Philosophical Review*, 110(3):361–396, 2001. doi: 10.1215/00318108-110-3-361.

W. Hodges. Compositional semantics for a language of imperfect information. *Logic Journal of the IGPL*, 5:539–563, 1997.

D. Ibeling and T. Icard. Probabilistic reasoning across the causal hierarchy. *arXiv preprint arXiv:2001.02889*, 2020.

S. Kaufmann. Conditional predictions. *Linguistics and Philosophy*, 28(2):181–231, 2005.

S. Kaufmann. Causal premise semantics. *Cognitive Science*, 37:1136–1170, 2013.

A. Kratzer. An investigation of the lumps of thought. *Linguistics and philosophy*, 12(5):607–653, 1989.

M. Lange. Laws, counterfactuals, stability, and degrees of lawhood. *Philosophy of Science*, 66(2):243–267, 1999.

D. Lewis. Counterfactuals and comparative possibility. In *Ifs*, pages 57–85. Springer, 1973.

D. Lewis. Causation. *The journal of philosophy*, 70(17):556–567, 1974.

D. Lewis. Counterfactual dependence and time's arrow. *NOÛS*, 13:455–476, 1979.

F. Liu. *Changing for the better: Preference dynamics and agent diversity*. PhD thesis, 2007.

M. Malink. *Aristotle's modal syllogistic*. Harvard University Press, 2013.

A. L. Mann, G. Sandu, and M. Sevenster. *Independence-Friendly logic - a game-theoretic approach*, volume 386 of *London Mathematical Society lecture note series*. Cambridge University Press, 2011. ISBN 978-0-521-14934-1.

J. Marti and R. Pinosio. Similarity orders from causal equations. In F. E. and J. Leite, editors, *Logics in Artificial Intelligence. JELIA 2014. Lecture Notes in Computer Science, vol 8761.*, pages 500–513. Springer, Cham, 2014a.

J. Marti and R. Pinosio. Similarity orders from causal equations. In *European Workshop on Logics in Artificial Intelligence*, pages 500–513. Springer, 2014b.

P. Nadathur and S. Lauer. Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa*, to appear.

S. Palan and C. Schitter. Prolific.ac — a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

R. Patterson. *Aristotle's modal logic: essence and entailment in the Organon*. Cambridge University Press, 2002.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

J. Pearl. *Causality. Models, reasoning, and inference*. Cambridge University Press, Cambridge, 2000.

J. Pearl. Causality: models, reasoning, and inference. *IIE Transactions*, 34(6): 583–589, 2002.

J. Pearl. *Causality. Models, reasoning, and inference*. Cambridge University Press, Cambridge, 2 edition, 2009.

J. Pearl. Structural counterfactuals: A brief introduction. *Cognitive Science*, 37: 977–85, 2013.

J. Pearl and D. Mackenzie. *The book of why. The new science of cause and effect*. Penguin Books UK, 2019.

J. Plaza. Logics of public communications. *Synthese*, 158(2):165–179, 2007.

N. Rescher. *Aristotle's theory of modal syllogisms and its interpretation*. University of Pittsburgh Press, 1964.

A. A. Rini et al. Is there a modal syllogistic? *Notre Dame Journal of Formal Logic*, 39(4):554–572, 1998.

L. J. Rips. Two causal theories of counterfactual conditionals. *Cognitive Science*, 34:175–221, 2010.

L. J. Rips and B. J. Edwards. Inference and explnation in counterfactal reasoning. *Cognitive Science*, 37:1107–1135, 2013.

P. Santorio. Interventions in premise semantics. *Philosophers' Imprint*, 19:1–27, 2019.

K. Schulz. "if you wiggle A, then B will change". causality and counterfactual conditionals. *Synthese*, 179(2):239–251, 2011.

K. Schulz. Minimal models vs. logic programming: the case of counterfactual conditionals. *Journal of Applied Non-Classical Logics*, 24(1-2):153–168, 2014.

K. Schulz and R. Rooij. Conditionals, causality and conditional probability. 2019.

K. Schulz and K. Xie. Revisiting king ludwig of bavaria: A causal account of epistemic counterfactuals. N.D.

K. Schulz, S. Smets, F. R. Velázquez-Quesada, and K. Xie. A logical and empirical study of right-nested counterfactuals. In *International Workshop on Logic, Rationality and Interaction*, pages 259–272. Springer, 2019.

B. Skyrms. The prior propensity account of subjunctive conditionals. In *Ifs*, pages 259–265. Springer, 1980.

S. A. Sloman and D. A. Lagnado. Do we "do"? *Cognitive Science*, 29:5–39, 2005.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 1st edition, 1993.

P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

R. C. Stalnaker. A theory of conditionals. In *Ifs*, pages 41–55. Springer, 1968.

W. Starr. Counterfactuals. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.

W. B. Starr. A uniform theory of conditionals. *Journal of Philosophical Logic*, 43 (6):1019–1064, 2014.

P. Thom. The two barbaras. *History and philosophy of logic*, 12(2):135–149, 1991.

J. Väänänen. *Dependence logic: a new approach to Independence Friendly logic*, volume 70 of *London Mathematical Society Student Texts*. Cambridge University Press, 2007.

J. van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2011. ISBN 978-0-521-76579-4.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library Series*. Springer, Dordrecht, The Netherlands, 2008. ISBN 978-1-4020-5838-7. doi: 10.1007/978-1-4020-5839-4.

H. P. Van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147(2):229–275, 2005.

J. van Eijck, M. Gattinger, and Y. Wang. Knowing values and public inspection. In *Indian Conference on Logic and Its Applications*, pages 77–90. Springer, 2017.

J. Van Rijen. *Aspects of Aristotle's logic of modalities*, volume 35. Springer Science & Business Media, 2012.

R. van Rooij and K. Schulz. A causal power semantics for generic sentences. *Topoi*, pages 1–16, 2019.

R. van Rooij and K. Xie. A causal analysis of modal syllogisms. N.D.

D. J. Vecchio. Essence and necessity, and the aristotelian modal syllogistic: A historical and analytical study. 2016.

F. Veltman. Making counterfactual assumptions. *Journal of Semantics*, 22(2): 159–180, 2005.

Y. Wang and Q. Cao. On axiomatizations of public announcement logic. *Synthese*, 190(1):103–134, 2013.

J. Woodward. *Making Things Happen*, volume 114 of *Oxford Studies in the Philosophy of Science*. Oxford University Press, 2003.

# Samenvatting

Dit proefschrift neemt de lezer mee op een intellectuele reis langs een reeks onderwerpen op het kruisvlak van het onderzoek naar conditionele zinnen, causaliteit en epistemologie. De focus gaat uit naar een aantal centrale problemen die op dit snijvlak liggen en die in de recente literatuur bijzonder veel aandacht hebben gekregen. Ik zal laten zien hoe, met behulp van een combinatie van kennis en gereedschap uit alle drie onderzoeksgebieden, wij substantieel vooruitgang kunnen boeken bij het oplossen van deze problemen. Ook zal ik duidelijk maken dat een aanpak die kennis uit alle drie gebieden integreert, ons een dieper inzicht geeft in de rijke connecties tussen conditionele zinnen, causaliteit en epistemologie.

Het centrale innovatief ingrediënt van mijn proefschrift ligt in de formele modellen die ik zal voorstellen om te komen tot een geïntegreerde representatie van causale en epistemische informatie. In deze modellen combineer ik de structureel functionele modellen, die gebruikt worden om causale verbanden formeel te representeren, met de (dynamisch) epistemische logica, een logica die epistemisch redeneren kan beschrijven. Het gebrek aan een goede expliciete weergave van epistemische informatie in causale modellen en adequate uitdrukkingen in de object-taal om over deze informatie te communiceren, is volgens mij op dit moment een groot struikelblok voor het onderzoek op het snijvlak tussen causaliteit, epistemologie en conditionele zinnen. Ik zal in het tweede en het derde hoofdstuk van mijn proefschrift twee verschillende modellen introduceren die causale en epistemische informatie combineren. Ik zal laten zien dat deze modellen ons helpen om centrale problemen uit de recente literatuur aan te pakken. Laat ik hierbij wel opmerken dat ik mijn resultaten slechts als een eerste stap zie, er is nog veel te doen om een volledige integratie van beide types informatie te bereiken. Ook op het gebied van toepassingen van deze modellen zijn er nog talloze mogelijkheden voor toekomstig onderzoek.

In de hoofdstukken 3 en 4 komen een aantal tekortkomingen van de zeer

populaire interventionistische aanpak voor counterfactuele zinnen aan bod. Deze causale benadering van hun betekenis is buitengewoon succesvol in het beschrijven van hoe wij counterfactuele zinnen begrijpen en ermee kunnen redeneren. Echter, de interventionistische aanpak kent ook een aantal problemen. Ten eerste kan deze theorie voor de betekenis van counterfactuele zinnen niet worden toegepast op voorbeelden die een epistemische lezing toelaten. Bouwend op resultaten uit het tweede hoofdstuk zal ik in hoofdstuk 3 laten zien dat een rijkere notie van een causaal model ons toestaat om de interventionistische aanpak uit te breiden naar dit type van zinnen. In dit nieuwe model staat opnieuw de weergave centraal van de epistemische toestand van de agent die de counterfactuele zin overweegt.

Hoofdstuk 4 richt zich op een ander probleem. De interventionistische benadering heeft ook moeite om de betekenis van verweven counterfactuele zinnen correct te beschrijven. In het bijzonder kijken wij in dit hoofdstuk naar counterfactuele zinnen die in hun consequent weer een counterfactuele zin hebben staan. In hoofdstuk 4 zal ik laten zien hoe we door een aanpassing te maken in de notie van interventie, die standaard wordt gebruikt, dit probleem kunnen oplossen. Het centrale idee achter deze nieuwe notie van interventie is dat ze geen wetten oplegt, maar slechts singuliere feiten van de actuele wereld aanpast.

Voor beide problemen die centraal staan in hoofdstuk 3 en 4 (de interpretatie van epistemische counterfactuele zinnen en de interpretatie van verweven counterfactuele zinnen) lukt het mij om een oplossing te presenteren waarbij we nog steeds het centrale onderliggende idee van interventionisme omarmen. Interventionisme blijft dus met haar hoofd boven het water. In het vijfde hoofdstuk van het proefschrift breng ik op een andere manier de kracht van de interventionistische benadering naar voren. Ik beargumenteer dat een dergelijke causale benadering van conditionele zinnen, nu in het bijzonder toegepast op indicatieve conditionele zinnen, ons ook kan helpen om Aristoteles' visie op modale syllogismen te begrijpen. Ik zal laten zien dat met behulp van zo'n benadering, een interpretatie van modale syllogismen mogelijk is die overeen komt met Aristoteles' oordeel over de geldigheid van deze redeneringen.

# Abstract

This dissertation is an intellectual journey along topics at the intersection of the study of conditionals, causality and epistemology. It will focus on a couple of problems at this intersection pointed out in recent research. I will demonstrate how by combining knowledge and tools from all three fields we can make substantial progress on solving these issues. I will also show that this integrated approach provides us with a better understanding of the relation between conditionals, causality and epistemology.

The most important innovation proposed in this thesis is the integration of structural functional models, used to represent causal dependencies, and (dynamic) epistemic logic used to represent the epistemic states of reasoning agents. The lack of a good explicit representation of epistemological information in causal models and the means to talk about this information in the object language is in my eyes a central limitation hindering progress in research at the intersection of causality, epistemology and conditionals. I will introduce two different versions of such integrated models, in chapter 2 and chapter 3, and show that they can be very helpful to answer open questions in the field. Though it needs to be said that this thesis only lays the ground work for such an integration between both formal models. Much more is possible and many questions still need to be answered in future work.

In the chapters 3 and 4 I will focus on limitations of the very popular intervetionist approach to counterfactuals. This causal approach to the meaning of these sentences has been shown to be very successful in capturing the way we interpret and reason with counterfactuals. However, the interventionist approach is also known to give rise to a number of mispredictions. First of all, this approach cannot handle examples in which a counterfactual seems to get an epistemic reading. Building on the results of chapter 2, we will show in chapter 3 that this problem can be addressed a richer notion of causal model that also allows us to represent the epistemic state of the agent engaged in counterfactual thinking.

Chapter 4 addresses a different problem: the interventionist approach also has a hard time accounting for right-nested counterfactuals, i.e. counterfactuals whose consequent contains another conditional sentence. In chapter 4 we will show that by modifying the notion of intervention used in the approach we can also deal with this problem. The central idea behind the new notion of intervention we propose is that instead of causal laws, it only changes particular facts in the world of evaluation.

Notice that in both cases, concerning the problem with epistemic counterfactuals and with regards to right-nested counterfactuals, I am able to overcome the limitations of the interventionist approach without giving up its central ideas. Thus even in these stormy waters the general approach perseveres. A different form of evidence for this approach to conditionals is provided in the fifth chapter of the thesis. Here I show that such a causal approach to conditionals, now extended to indicative conditionals, can help us to understand Aristoteles' view on modal syllogisms. I will demonstrate that with such an approach at hand one can give an interpretation of modal syllogisms that confirms Aristotle's validity judgements.

*Titles in the ILLC Dissertation Series:*

ILLC DS-2009-01: **Jakub Szymanik**
*Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language*

ILLC DS-2009-02: **Hartmut Fitz**
*Neural Syntax*

ILLC DS-2009-03: **Brian Thomas Semmes**
*A Game for the Borel Functions*

ILLC DS-2009-04: **Sara L. Uckelman**
*Modalities in Medieval Logic*

ILLC DS-2009-05: **Andreas Witzel**
*Knowledge and Games: Theory and Implementation*

ILLC DS-2009-06: **Chantal Bax**
*Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.*

ILLC DS-2009-07: **Kata Balogh**
*Theme with Variations. A Context-based Analysis of Focus*

ILLC DS-2009-08: **Tomohiro Hoshi**
*Epistemic Dynamics and Protocol Information*

ILLC DS-2009-09: **Olivia Ladinig**
*Temporal expectations and their violations*

ILLC DS-2009-10: **Tikitu de Jager**
*"Now that you mention it, I wonder...": Awareness, Attention, Assumption*

ILLC DS-2009-11: **Michael Franke**
*Signal to Act: Game Theory in Pragmatics*

ILLC DS-2009-12: **Joel Uckelman**
*More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains*

ILLC DS-2009-13: **Stefan Bold**
*Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.*

ILLC DS-2010-01: **Reut Tsarfaty**
*Relational-Realizational Parsing*

ILLC DS-2010-02: **Jonathan Zvesper**
*Playing with Information*

ILLC DS-2010-03: **Cédric Dégremont**
*The Temporal Mind. Observations on the logic of belief change in interactive systems*

ILLC DS-2010-04: **Daisuke Ikegami**
*Games in Set Theory and Logic*

ILLC DS-2010-05: **Jarmo Kontinen**
*Coherence and Complexity in Fragments of Dependence Logic*

ILLC DS-2010-06: **Yanjing Wang**
*Epistemic Modelling and Protocol Dynamics*

ILLC DS-2010-07: **Marc Staudacher**
*Use theories of meaning between conventions and social norms*

ILLC DS-2010-08: **Amélie Gheerbrant**
*Fixed-Point Logics on Trees*

ILLC DS-2010-09: **Gaëlle Fontaine**
*Modal Fixpoint Logic: Some Model Theoretic Questions*

ILLC DS-2010-10: **Jacob Vosmaer**
*Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.*

ILLC DS-2010-11: **Nina Gierasimczuk**
*Knowing One's Limits. Logical Analysis of Inductive Inference*

ILLC DS-2010-12: **Martin Mose Bentzen**
*Stit, Iit, and Deontic Logic for Action Types*

ILLC DS-2011-01: **Wouter M. Koolen**
*Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice*

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
*Small steps in dynamics of information*

ILLC DS-2011-03: **Marijn Koolen**
*The Meaning of Structure: the Value of Link Evidence for Information Retrieval*

ILLC DS-2011-04: **Junte Zhang**
*System Evaluation of Archival Description and Access*

ILLC DS-2011-05: **Lauri Keskinen**
*Characterizing All Models in Infinite Cardinalities*

ILLC DS-2011-06: **Rianne Kaptein**
*Effective Focused Retrieval by Exploiting Query Context and Document Structure*

ILLC DS-2011-07: **Jop Briët**
*Grothendieck Inequalities, Nonlocal Games and Optimization*

ILLC DS-2011-08: **Stefan Minica**
*Dynamic Logic of Questions*

ILLC DS-2011-09: **Raul Andres Leal**
*Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications*

ILLC DS-2011-10: **Lena Kurzen**
*Complexity in Interaction*

ILLC DS-2011-11: **Gideon Borensztajn**
*The neural basis of structure in language*

ILLC DS-2012-01: **Federico Sangati**
*Decomposing and Regenerating Syntactic Trees*

ILLC DS-2012-02: **Markos Mylonakis**
*Learning the Latent Structure of Translation*

ILLC DS-2012-03: **Edgar José Andrade Lotero**
*Models of Language: Towards a practice-based account of information in natural language*

ILLC DS-2012-04: **Yurii Khomskii**
*Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.*

ILLC DS-2012-05: **David García Soriano**
*Query-Efficient Computation in Property Testing and Learning Theory*

ILLC DS-2012-06: **Dimitris Gakis**
*Contextual Metaphilosophy - The Case of Wittgenstein*

ILLC DS-2012-07: **Pietro Galliani**
*The Dynamics of Imperfect Information*

ILLC DS-2012-08: **Umberto Grandi**
*Binary Aggregation with Integrity Constraints*

ILLC DS-2012-09: **Wesley Halcrow Holliday**
*Knowing What Follows: Epistemic Closure and Epistemic Logic*

ILLC DS-2012-10: **Jeremy Meyers**
*Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies*

ILLC DS-2012-11: **Floor Sietsma**
*Logics of Communication and Knowledge*

ILLC DS-2012-12: **Joris Dormans**
*Engineering emergence: applied theory for game design*

ILLC DS-2013-01: **Simon Pauw**
*Size Matters: Grounding Quantifiers in Spatial Perception*

ILLC DS-2013-02: **Virginie Fiutek**
*Playing with Knowledge and Belief*

ILLC DS-2013-03: **Giannicola Scarpa**
*Quantum entanglement in non-local games, graph parameters and zero-error information theory*

ILLC DS-2014-01: **Machiel Keestra**
*Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms*

ILLC DS-2014-02: **Thomas Icard**
*The Algorithmic Mind: A Study of Inference in Action*

ILLC DS-2014-03: **Harald A. Bastiaanse**
*Very, Many, Small, Penguins*

ILLC DS-2014-04: **Ben Rodenhäuser**
*A Matter of Trust: Dynamic Attitudes in Epistemic Logic*

ILLC DS-2015-01: **María Inés Crespo**
*Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.*

ILLC DS-2015-02: **Mathias Winther Madsen**
*The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science*

ILLC DS-2015-03: **Shengyang Zhong**
*Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory*

ILLC DS-2015-04: **Sumit Sourabh**
*Correspondence and Canonicity in Non-Classical Logic*

ILLC DS-2015-05: **Facundo Carreiro**
*Fragments of Fixpoint Logics: Automata and Expressiveness*

ILLC DS-2016-01: **Ivano A. Ciardelli**
*Questions in Logic*

ILLC DS-2016-02: **Zoé Christoff**
*Dynamic Logics of Networks: Information Flow and the Spread of Opinion*

ILLC DS-2016-03: **Fleur Leonie Bouwer**
*What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm*

ILLC DS-2016-04: **Johannes Marti**
*Interpreting Linguistic Behavior with Possible World Models*

ILLC DS-2016-05: **Phong Lê**
*Learning Vector Representations for Sentences - The Recursive Deep Learning Approach*

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
*Aligning the Foundations of Hierarchical Statistical Machine Translation*

ILLC DS-2016-07: **Andreas van Cranenburgh**
*Rich Statistical Parsing and Literary Language*

ILLC DS-2016-08: **Florian Speelman**
*Position-based Quantum Cryptography and Catalytic Computation*

ILLC DS-2016-09: **Teresa Piovesan**
*Quantum entanglement: insights via graph parameters and conic optimization*

ILLC DS-2016-10: **Paula Henk**
*Nonstandard Provability for Peano Arithmetic. A Modal Perspective*

ILLC DS-2017-01: **Paolo Galeazzi**
*Play Without Regret*

ILLC DS-2017-02: **Riccardo Pinosio**
*The Logic of Kant's Temporal Continuum*

ILLC DS-2017-03: **Matthijs Westera**
*Exhaustivity and intonation: a unified theory*

ILLC DS-2017-04: **Giovanni Cinà**
*Categories for the working modal logician*

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
*Communication and Computation: New Questions About Compositionality*

ILLC DS-2017-06: **Peter Hawke**
*The Problem of Epistemic Relevance*

ILLC DS-2017-07: **Aybüke Özgün**
*Evidence in Epistemic Logic: A Topological Perspective*

ILLC DS-2017-08: **Raquel Garrido Alhama**
*Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence*

ILLC DS-2017-09: **Miloš Stanojević**
*Permutation Forests for Modeling Word Order in Machine Translation*

ILLC DS-2018-01: **Berit Janssen**
*Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs*

ILLC DS-2018-02: **Hugo Huurdeman**
*Supporting the Complex Dynamics of the Information Seeking Process*

ILLC DS-2018-03: **Corina Koolen**
*Reading beyond the female: The relationship between perception of author gender and literary quality*

ILLC DS-2018-04: **Jelle Bruineberg**
*Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems*

ILLC DS-2018-05: **Joachim Daiber**
*Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation*

ILLC DS-2018-06: **Thomas Brochhagen**
*Signaling under Uncertainty*

ILLC DS-2018-07: **Julian Schlöder**
*Assertion and Rejection*

ILLC DS-2018-08: **Srinivasan Arunachalam**
*Quantum Algorithms and Learning Theory*

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
*Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks*

ILLC DS-2018-10: **Chenwei Shi**
*Reason to Believe*

ILLC DS-2018-11: **Malvin Gattinger**
*New Directions in Model Checking Dynamic Epistemic Logic*

ILLC DS-2018-12: **Julia Ilin**
*Filtration Revisited: Lattices of Stable Non-Classical Logics*

ILLC DS-2018-13: **Jeroen Zuiddam**
*Algebraic complexity, asymptotic spectra and entanglement polytopes*

ILLC DS-2019-01: **Carlos Vaquero**
*What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance*

ILLC DS-2019-02: **Jort Bergfeld**
*Quantum logics for expressing and proving the correctness of quantum programs*

ILLC DS-2019-03: **András Gilyén**
*Quantum Singular Value Transformation & Its Algorithmic Applications*

ILLC DS-2019-04: **Lorenzo Galeotti**
*The theory of the generalised real numbers and other topics in logic*

ILLC DS-2019-05: **Nadine Theiler**
*Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles*

ILLC DS-2019-06: **Peter T.S. van der Gulik**
*Considerations in Evolutionary Biochemistry*

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
*Cuts and Completions: Algebraic aspects of structural proof theory*

ILLC DS-2020-01: **Mostafa Dehghani**
*Learning with Imperfect Supervision for Language Understanding*

ILLC DS-2020-02: **Koen Groenland**
*Quantum protocols for few-qubit devices*

ILLC DS-2020-03: **Jouke Witteveen**
*Parameterized Analysis of Complexity*

ILLC DS-2020-04: **Joran van Apeldoorn**
*A Quantum View on Convex Optimization*

ILLC DS-2020-05: **Tom Bannink**
*Quantum and stochastic processes*