TU WIEN Informatics

A comprehensive Survey of the Actual Causality Literature

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Logic and Computation

eingereicht von

Konstantin Raphael Kueffner, B.Sc.
Matrikelnummer 01252260

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dr. Agata Ciabattoni

Wien, 8. Oktober 2021

Konstantin Raphael Kueffner	Agata Ciabattoni

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Informatics

A comprehensive Survey of the Actual Causality Literature

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Logic and Computation

by

Konstantin Raphael Kueffner, B.Sc.
Registration Number 01252260

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dr. Agata Ciabattoni

Vienna, 8th October, 2021

Konstantin Raphael Kueffner | Agata Ciabattoni

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Erklärung zur Verfassung der Arbeit

Konstantin Raphael Kueffner, B.Sc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 8. Oktober 2021

Konstantin Raphael Kueffner

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Danksagung

Zunächst möchte ich mich bei meiner Betreuerin Univ.Prof. Dr. Agata Ciabattoni für Ihr fast unmenschlich schnelles Feedback, Ihre Unterstützung und Ihren Rat ausdrücklichst bedanken. Außerdem möchte ich mich dafür bedanken, dass Sie es mir ermöglicht hat dieses Thema frei zu erkunden. Vor allem bin ich dankbar für die immense Geduld die Sie mir während des Projektes gezeigt hat.

Zweitens möchte ich Univ.Prof. Dr. Hans Tompits, Prof. Dr. Hans Göpfrich, Prof. Dr. Stefan Sobernig und Prof. Dr. Mark Strembeck für die jeweiligen Anstellungen bedanken. Diese haben es mir ermöglicht verschiedene Aspekte des wissenschaftlichen Arbeitsumfelds näher kennenzulernen. Besonderer Dank geht and den Herrn Prof. Dr. Mark Strembeck für sein Vertrauen, seine Toleranz und all die Möglichkeiten die er mir gewährt hat.

Drittens möchte ich O.Univ.Prof. Dr. Thomas Eiter, Univ.Prof. Dr. Laura Kovacs, Prof. Dr. Stefan Sobernig, Ao.Univ.Prof. Dr. Christian Fermüller, und Sonja Morzycki für deren direkte Hilfe bei der Sicherung meiner aktuellen Doktorandenstelle am IST Austria danken.

Viertens möchte ich meinen Eltern, Großeltern, meiner Tante und meinem Bruder für all ihre Liebe und Unterstützung danken. Danke an meinen Großvater, der mir das Studieren ermöglichte. Danke an meine Tante für Ihren Rat. Danke an meine Eltern, für ihr volles Vertrauen in meine Entscheidungen. Danke an meine Großmütter, für ihre Fürsorge. Danke an meinen Bruder, für alles. Zudem vielen Dank an alle meine Freunde, die mir geholfen haben die Volatilität der vergangenen Jahre gut zu überstehen.

Zum Schluss, meinen Dank an Sonja Morzycki und Gerald Kimmersdorfer für das Korrekturlesen der Danksagungen und der deutschen Kurzfassung.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

Firstly, I would like to thank my advisor Univ.Prof. Dr. Agata Ciabattoni for her almost inhumanly quick feedback, guidance and overall support. Moreover, I would like to thank her for allowing me to freely explore this topic. Most importantly I am grateful for the immense patience she displayed while supervising me.

Second, I would like to thank Univ.Prof. Dr. Hans Tompits, Prof. Dr. Hans Göpfrich, Prof. Dr. Stefan Sobernig and Prof. Dr. Mark Strembeck for employing me and providing me with insights into the academic work environment. In particular, I have to thank Prof. Dr. Mark Strembeck for his trust, tolerance and the opportunities he granted me with.

Third, I would like to thank O.Univ.Prof. Dr. Thomas Eiter, Univ.Prof. Dr. Laura Kovacs, Prof. Dr. Stefan Sobernig, Ao.Univ.Prof. Dr. Christian Fermüller, and Sonja Morzycki for their help in securing my current PhD-position at IST Austria.

Fourth, I would like to thank my parents, grandparents, aunt and brother for all their love and support. In particular, I would like to thank my grandfather for providing me with the opportunity to study, my aunt for advice, my parents for their immense trust, my grandmothers for their care and my brother for everything. Moreover, many thanks to all my friends, who helped me to deal with the volatility of the past years.

Lastly, thanks to Sonja Morzycki and Gerald Kimmersdorfer who proofread the acknowledgements and German version of the abstract respectively.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Der Forschungsbereich um Kausalität gewann in der Informatik in den letzten Jahren zunehmend an Bedeutung. Eine Definition die Kausalität formal erfasst würde es Computern ermöglichen, "Warum"-Fragen zu beantworten und hätte vielversprechende Anwendungen im Bereich der Verifikation, des maschinellem Lernen, der Erklärbarkeit, dem formalen rechtlichem Schließen und der algorithmischer Gerechtigkeit führen. Um das Erreichen zu können müssen kausale Beziehungen aus Daten abgeleitet werden. Diese werden anschließend verwendet um die tatsächlichen Ursachen für Ereignisse in konkreten Situationen zu identifizieren. Das Ermitteln solcher Ursachen wird als token-kausale Inferenz bezeichnet. Nach jetzigem Kenntnisstand existiert kein ausreichend umfangreiches Werk, welches den aktuellen Stand der Technik im Bereich token-kausaler Inferenzsysteme offenlegt. Diese Dissertation soll das eben genannte Defizit begleichen. Die hierfür durchgeführte Literaturrecherche ist in drei verschiedene Granularitätsebenen unterteilt. Die erste Ebene betrachtet die Literatur als eigenständiges Studienobjekt. Im Kontext dessen werden Techniken der Netzwerkanalyse verwendet, um wichtige Publikationen, Autoren und Forschungsgemeinschaften zu identifizieren. Die zweite Ebene ist eine klassische Literaturrecherche, bei der eine Teilmenge der gesammelten Literatur im Detail untersucht wird. Das Ziel hierbei ist es die wichtigsten Werkzeuge zur Formalisierung von Kausalität zu extrahieren, zu beschreiben und zu kategorisieren. Dieser Teilmenge gehören unter anderem die formalen Sprachen zur Codierung kausaler Beziehungen an, aber auch die verschiedenen Kausalitätsdefinitionen sowie diverse Szenarien welche verwendet werden um diese Definitionen zu testen. Die dritte Ebene beschäftigt sich mit vier solcher Kausalitätsdefinitionen im Detail. Diese werden formal eingeführt und anhand der vorgestellten Testszenarien verglichen. Dieser letzte Teil erforderte einige Originalarbeiten, da nicht alle Szenarien in der Literatur formalisiert zu finden sind.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Abstract

The study of causality has recently gained traction in computer science. Formally capturing causal reasoning would allow computers to answer "Why"-questions and would result in significant advances in fields such as verification, machine learning, explainability, legal reasoning and algorithmic fairness. To accomplish this, one needs to be able to infer type causal relationships, i.e. general statements about causal dependencies, from data and then use those relationships to identify the actual causes of an event in a given situation; such causes are referred to as token causes. To the best of our knowledge, there does not exist a comprehensive survey, reviewing the state of the art of formal systems for token causality. The present thesis addresses this deficit. The literature review that we have performed operates on three different levels of granularity. The first considered the literature landscape itself as an object of study, employing network analysis techniques to identify important publications, authors and research communities. The second is a classical literature review, where a subset of the collected literature is investigated in detail, to extract, describe and categorise the tools used for formalising causation. This includes the languages for encoding causal relationships, the various definitions that try to capture token causality, as well as the benchmark used to test the capabilities of those definitions. In the third part we describe and compare the four main token causality definitions, w.r.t. the most prominent benchmarks in the literature. This last part also required some original work, as not all the examples are found in the literature.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Kurzfassung **xi**

Abstract **xiii**

Contents **xv**

1 Introduction **1**

2 Literature Collection and Analysis **7**
2.1 Methodology 8
2.2 Analysis 16

3 Formalising Causation: A Survey **37**
3.1 Token Causality: Languages 38
3.2 Token Causality: Definitions 52
3.3 Token Causality: Benchmarks 63

4 Comparing recent systems for token causality **87**
4.1 Definitions 87
4.2 Benchmarks 107

5 Conclusion **137**

6 Appendix **143**

List of Figures **153**

List of Tables **155**

List of Algorithms **157**

Bibliography **159**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

Causal inference is a seemingly integral part of human reasoning [Rei13, p. 733–752] it enables us to answer "Why?" questions, i.e. it provides humans with the ability to formulate explanations. [PM18]. Naturally, a topic of such magnitude has captivated philosophers for millennia, including notable figures such as Aristotle, Hume, and Kant, as well as some members of the Vienna Circle [BHM09, p. 21,73,92,108].

In recent years the study of causality has departed from being a solely philosophical exercise and gained traction across a multitude of fields. One of which is computer science, which now harbours various communities each trying to approach causality from different formal perspectives.

The causality literature that intersects with computer science can roughly be separated into two sub-fields. The first is concerned with deriving type causal statements, i.e. general statements about causal relationships. The second presupposes a set of type causal relations and tries to identify the actual causes of an event in a given situation. This type of causation is known as token causality or actual causality. The former naturally intersects with the domain of machine learning, where the objective is to learn patter from data. Token causality is instead closely related to those areas of computer science concerned with formal reasoning and the goal of this sub-field is to axiomatise the intuitive form of causal reasoning used by humans. This is an important task as a sufficiently robust formalisation of token causality could allow a computer to explain themselves. While this would impact many areas of computer science, including machine learning, verification, and data bases, recent political developments[1] emphasise the need for such a project.

Although the roots are old, it was Judea Pearl who recognised the need for causal reasoning in computer science. With the development of his causal calculus, he popularised the

[1] see "The Right to Explanation" [SP18]

study of causality in the context of computer science, so much so that applications of causal reasoning can be found across many of the major areas in computer science. For example, the study of causality intersects with the logic literature, e.g. [MT+97, Boc03], it appears in the context of logic programming e.g. [GLL+04, LY10] and even has found application in the area of program verification and databases [Kup16, Hal16a, p. 205-211]. Moreover, being heavily influenced by Pearl, causality can be found in classical machine learning literature such as probabilistic graphical models, e.g. [Pea09, Sch19]. Furthermore, recent voices in the neural network community frequently appeal to the necessity of combining causal reasoning and neural networks [BDR+19]. Lastly, mentions of causality permeate well into the outskirts of computer science, including areas such as formal legal reasoning [LSW19b] as well as algorithmic fairness [KLRS17].

The examples of causal reasoning mentioned above refer to causation in general, which to no surprise is dominated by the ideas of Judea Pearl. The literature discussing token causality is dominated instead by the ideas of Joseph Halpern, who has introduced several definitions of token causality [Hal16a], all of which use counterfactuals as their core mechanic. Other common approaches for defining token causality are default reasoning [Boc18a], processes [BS18] and probability theory [Ven11]. A hallmark of this literature is that it is plagued by disagreement on every level, i.e. the general approach, the philosophical tradition, how to formalise benchmarks, how to interpret benchmarks, and so on. Therefore, it is hard to find a single consensus, that can be used to measure how close we are to formally capturing the elusive concept of token causality. Even more troublesome, it is not even clear whether such a formalism can exists.

Given this state of affairs, it comes to no surprise that as of now the dust has not yet settled on a single formalism that "correctly" captures token causality. Therefore, the literature being littered with competing definitions, it is rather difficult to assess the current state of the art. This also happens because most of the publications in this area discuss no more than two definitions at a time. To rectify this issue, this thesis is a systematic literature survey collecting and contrasting all the various attempts of formalising the elusive concept of token causality. To the best of our knowledge, this is the first comprehensive survey of token causal reasoning definitions. The only publication similar to this thesis is [Wes15], which contrasts several formalisms at once, most of which were introduced more than a decade ago.

Given the vast body of literature surrounding causation, this survey operates on three different levels of granularity.

The first level, discussed in Chapter 2, is a structural literature survey. It analyses the literature based on features such as citation and co-authorship relations. The objective of this part is to answer questions such as, what are the important publications; which authors wield the most influence; which authors work on similar subjects. Meaning that it intentionally avoids investigations into the subject matters discussed in the literature, and approaches the literature landscape as its own object of study. This is accomplished by casting the relationships between publications into a series of new graphs. To be more precise, this part of the survey utilises a snowball search strategy for collecting the required

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

publications. That is, the start set will consist of all publications from the "Journal Knowledge-Bases Systems", the "Journal Artificial Intelligence", the "Journal Artificial Intelligence and Law" and "International Joint Conferences on Artificial Intelligence Organization" that were published between 01.2017 and 3.2020, which amounts to 4223 unique publications. Those are reduced by keyword search to a manageable set of 37 publications. From there, several forward-snowball, backward-snowball, and filter steps are performed, resulting in a total of 872 unique publications, which are superficially reviewed for relevancy, resulting in 294 publications. After filtering by publication date, the resulting set of 107 publications is placed into a citation (and co-authorship) graph and analysed using some network theoretic measures, e.g. centrality measures, clustering, and others. This analysis produces a set of 36 important publications, which is passed down to the second level. This part of the survey is partially automated, utilising web crawlers and PDF-parsers to fill the SQL-Light Database, which is subsequently analysed using the iGraph package [CN+06].

The second level of this survey, discussed in Chapter 3, is a classical literature review of the previously established set of important publications. The objective of this part is to provide an overview of the available formal languages, the definitions, and the benchmarks used for token causal reasoning. This includes an intuitive introduction of each language, definition, and benchmark, as well as their categorisation based on dimensions such as time, popularity, and other characteristics. This part of the survey builds on the 36 important publications identified in Chapter 2, by extracting each formal language, token causal definition, or mentioned benchmark. This process identified 18 unique languages, 32 unique definitions of token causality, and more than 20 benchmarks. Using the set of important publications the first step is to assess the popularity of each construct. Each of the collected constructs are then categorised and surveyed independently. Apart from a short discussion, this includes a categorisation of languages and definitions. Languages are categorised based on properties such as quantification, many-valued variables, default reasoning, temporal reasoning, or probabilistic reasoning. Definitions will be categorised based on their language, and whether they follow a counterfactual, process orientated, probabilistic, or regularity theoretic approach.

The third level, discussed in Chapter 4, is an in-depth introduction of selected few definitions and languages. This includes, a definition provided by Halpern, which is formulated using causal models [Hal15a]; a definition introduced in [BV18] which utilises a slightly modified version of causal models; a definition provided by [DBV19] which uses a version of CP-Logic; a definition developed in [Boc18a] which builds on the Non-Monotonic Theory presented in [MT+97]. The objective of this part is to provide insights into the various kinds of machinery used for token causal inference. This includes a formal introduction of the languages and definitions, as well as a comparison of the approaches based on common benchmarks.

We start by providing a rudimentary discussion on the differences between token and type causality and briefly presents an overview of the main philosophical traditions used to define causality. The most fundamental distinction made in the causality

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

literature is the distinction between type and token causality, which is often referred to as actual causality [Hal16a]. However, there are many more approaches to how one can conceptualise causality. For example, according to [BHM09] there are standard and alternative approaches to causation. The standard approaches to causation include regularity theories, counterfactual theories, probabilistic theories, causal process theories and agency interventionist theories, while the latter includes theories about causal power and capacities, an anti-reductionist approach, the field of causal modelling, an approach requiring the existence of causal mechanism and one that embraces pluralism. To provide a basic understanding of causality, this section elaborates on the notion of type and token causality, as well as the standard approaches to causation.

The classification of causality into type and token causality is rooted in the metaphysical distinction between types and tokens, which is used to differentiate a general sort of thing and its particular occurrence [Wet18]. [Hau05] use the statement "Rose is a rose is a rose is a rose." to highlight the differences between types and tokens. That is, depending on what one may understand as word, this sentence contains three or ten different words. In the prior, the word-types of the sentence are counted, while in the latter the word-tokens are counted.

Similarly, one can distinguish two (possibly distinct) notions of causality. Type causality is concerned with forward-looking statements such as "smoking causes lung cancer", granting their wielder some predictive capabilities. Hence, establishing type causality is often the pursuit of scientific enquiry. However, this suggests that type-causal relations do not establish a strong causal connection, but rather a causal tendency. Meaning, while smoking may cause lung cancer, it is not necessary the case that a smoker will develop lung cancer, thus the statement "smoking tends cause lung cancer" may be more precise. By contrast, if one wants to establish that the act of smoking caused lung cancer in a particular person, one speaks of token causality. The objective hereby is to identify the events that explains why a certain outcome arose. Hence, token causality tends to be backwards-looking. Unfortunately, there remains debate about whether those two notion of causation are distinct. For example, one view to take is that type causation is merely a generalisation of token causal relations, which are assumed to be fundamental. Another view would be to assume that token causation is merely an instantiation of type-level laws, which under this perspective are considered as the fundamental element. Yet another view considers both type and token causality to be distinct expressions of a singular unknown causal relation. For example, Halpern's token causal inference mechanism requires a model of the world which is given as a set of equations that encode type causal relations [Hau05, Hal16a].

The debate of what constitutes token or type causality can be extended to variables. That is, are causal relationships established between variables or the values of those variables. In the former case, the relations would be considered type-level relations and in the latter case, the relations would be considered token-level relations [Hau05]. For example, consider a simplistic model of causality. That is, we are given two variables X and Y, lets say X is a type-level cause of Y if there exists a possible intervention on the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

variable X such that the value of the variable Y changes. By contrast, the value x of X is a token cause for the value y of Y, if the value x is essential for the fact that variable Y has value y [Wes15].

While the distinction of type and token causality addresses the phenomena of causality directly, the five standard approaches found in [BHM09], i.e. regularity theories, counterfactual theories, probabilistic theories, causal process theories and agency interventionist theories, can be used to classify definitions of token causality based on their philosophical foundations.

Both regularity and counterfactual theories of causation seem to be the most commonly discussed in the computer science literature. The regularity theoretic view on causation defines causality by utilising regularities and is therefore strongly connected to the notion of type causality, e.g. an event A is a cause of event B, if A usually precedes B. Unfortunately, the simplistic view sketched here is clearly insufficient, e.g. as otherwise the night would be caused by the day. However, there are more refined versions of this principle that closer approximate causation, e.g. [Bau13] and [Boc18a]. By contrast, the counterfactual theoretic view on causation relies on hypothetical statements, e.g. an event A is a cause of event B, if event A had not occurred then event B would not have occurred. Again this simplistic view of counterfactuals has proven to be insufficient as well, because there are scenarios where there is no direct counterfactual dependence between cause and effect, e.g. two people pushing the same button at the same time. Nevertheless, many modern definitions of causality adhere to this framework, the most prominent of which are the definitions produced by Halpern, see [HP05] and [Hal15a]. Given their differences, it is quite interesting that both can trace their origin to Hume [Hum48, Hal16a, p. 2].

> We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or in other words, where, if the first object had not been, the second never had existed.

The other approaches seem to be less prominent. Probabilistic theories adhere to the view that, if event A is the cause of event B, then the occurrence of event A increases the probability of event B occurring. Due to the probabilistic nature of the relationship between cause and effect, Bayes nets could be considered as one of the formalisms rooted in this theory. A major issue with this approach is that there may be probabilistic relationships between variables, which are not necessarily causal ones, and teasing them apart provides a major challenge for this tradition. Causal process theories, distinguish themselves, by trying to characterise causation using continuous processes and the relationships between them, rather than trying to find a suitable relation between discrete events. Lastly, agency interventionist theories, seem to be related to counterfactual theories of causation, but differentiate themselves by requiring that the hypotheticals generated using interventions are tied to human agency [BHM09].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

To summarise, the structure of the thesis is a follows. Chapter 2 investigates the structure of the causality literature landscape using citation graphs to identify important publications and influential authors. Chapter 3 surveys a subset of important publications to summarise and categorise the various approaches taken to define token causality. Chapter 4 highlights selected formalisms by demonstrating their mechanism and by comparing them using common benchmarks. This includes definitions such as the one developed in [BV18] which relies on causal models; the one introduced in [DBV19] which uses CP-Logic; the one discussed in [Boc18a] which builds on Non-Monotonic Theory; the newest one put forward by Halpern, which was introduced in [Hal15a]. Moreover, we use the benchmarks referred to in the literature under the names "Symmetric Overdetermination", "Switching", "Late Preemption", "Early Preemption", "Double Preemption", "Bogus Preemption" and "Short-Circuiting". Chapter 5 concludes this thesis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

Literature Collection and Analysis

Intersecting with a wide range of subjects, e.g. philosophy, statistics, computer science, law, natural science, social sciences and more, as well as stretching over centuries, e.g. being already discussed by Hume in the 18th century, inquiries into causality are have produced an incredible the wealth of literature. Hence, to remain within a reasonable scope, it is of utmost importance to rely on a properly defined methodology and a suitable set heuristics, to navigate this vast ocean of literature. Therefore, the primary objective of this chapter is to outline the methodology employed in the collection of the literature used for the survey found in Chapter 3, where we will identify important formal languages used for encoding causal relationships, important token causal definitions, as well as important benchmarks developed for testing said definitions. Hence, Section 2.1 limits itself to a detailed characterisation of the publication collection process, as well as the introduction and justification of the methods used to identify relevant literature and influential authors among the collected publications.

It's secondary objective, mostly the subject of Section 2.2, is to answer questions such as, what are the important publications; which authors wield the most influence; which authors work on similar subjects. Meaning that it intentionally avoids investigations into the subject matters discussed in literature. This is accomplished by extracting information from the citation and co-authorship relations in the collected literature and using it to generate a series of pointers to potentially relevant publications, authors and research communities. Thereby, drawing a map of the causation literature landscape that highlights promising points of entry, which hopefully supports the reader in their voyage through the causation literature. By doing so this thesis is essentially treating the literature landscape surrounding causations as an object of study in itself.

2.1 Methodology

This section provides a detailed description of the methodology used to identify, collect and analyse the computer science, logic and philosophy literature surrounding causality. Firstly, data collection. The data, i.e. publications, are collected using a snowball search strategy, which is a search approach for systematic literature studies and refers to the use of reference list of a publications or the citations to those publications to identify additional publications [Woh14]. An important part of such an approach is a detailed characterisation of the set of publications from which snowball steps are conducted, as well as an adequate description of how and when those steps are employed. This information is provided in Subsection 2.1.1. Moreover, this subsection provides a detailed description of the publicly available database[1] used to store the meta information of the collected publications, the purpose of which is to provide other researches access to the constructed snapshot of the literature and to enhance transparency. Lastly, the methods used for the analysis of the collected data are discussed in Subsection 2.1.2.

2.1.1 Data Collection

The methodology underlying this systematic literature review employs a snowball search strategy. In general, according to [Woh14] any snowball search strategy should start by characterising an appropriate initial set of publications, i.e. the start set, which is then iteratively expanded by either forward or backward-snowballing until a desirable final set of publications is obtained. The start set should satisfy the following criteria:

- The start set should cover a diversity of communities.
- The number of papers in the start set should not be too small.
- The number of papers in the start set should not be too big.
- The start set should cover several different publishers, years and authors.
- The start set ought to be formulated from keywords (and their synonyms) in the research question.

Moreover, as stated in [Woh14] any snowball step on a given set of publications consists of both forward and backward-snowballing. The latter, adds all relevant references from all unprocessed publications to the set of publications. By contrast, the former leverages modern technologies, such as Google Scholar to identify every relevant publication that references any unprocessed publication in the provided set [Woh14].

Using this as a template, the actual methodology is constructed as follows. Firstly, the objectives that ought to be satisfied by the snowball search strategy are made explicit. In this particular case those objectives are to

[1] https://github.com/KonstantinRK/CausalitySurvey

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- focus on token causality publications;
- focus primarily on publications related to computer science, and artificial intelligence in particular;
- focus secondarily on publications related to philosophy or law;
- focus on publication that approach causality with sufficient formality;
- focus on logic and rule based approaches to causality;
- focus on the recent literature, i.e. publications between 2010 and (early) 2020.

Being a snowball search, the growth rate of the publications to consider is exponential. Hence, to serve the outlined objectives it is vital to construct a starting set that provides a sufficient strong directive. Since the primary focus is to remain within the greater context of computer science, logic and (symbolic) artificial intelligence, the start set construction is initiated by considering all articles from

- Journal Knowledge-Bases Systems (KBS)
- Journal Artificial Intelligence (AI)
- Journal Artificial Intelligence and Law (AI&Law)
- International Joint Conferences on Artificial Intelligence Organization (IJCAI)

that were published between 01.2017 and 3.2020. Focusing on such recent publications should serve the recency bias established in the methodology's objectives. The collected publications are subsequently preprocessed using a simple keyword search. That is the first necessary condition for a publication to be in the start set is

- that its title contains a string starting with the character sequence *"caus"* or
- its abstract contains a string starting with the character sequence *"causal"*.

Let $\mathcal{S}_0$ be the subset of all collected publications, that satisfy these criteria. To focus on logic and rule-based approaches, all publications that are deemed irrelevant under closer inspection or are inaccessible will be removed. The classification as relevant is done based on a list of soft criteria. By satisfying positive criteria the publication increases its chance of being deemed relevant, satisfying negative ones decreases its chance, and criteria marked by "$*$" are necessary.

$*$ Does the publication discuss causality or any related concepts?

$+$ Does the publication engage with the philosophical aspects of causality?

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

+ Does the publication try to formalise causality using logic (or another formal language)?
+ Does the publication's title explicitly mention logic and/or causality?
+ Does the publication discuss token causality?
– Does the publication discuss causality in the context of machine learning?
– Does the publication discuss causality in a highly informal manner?
– Is the publication a book?

To explain the snowballing step, some general notation must be introduced. Let $\mathcal{X}$ be some set of publications. Then $\mathcal{X}^c$ is the set of publications deemed relevant by the previously stated criteria. Furthermore, let $\mathcal{X}^r$ be the set of publication deemed relevant by the previously stated criteria, which are published after (and including) 2010.

Utilising this notation, let $\mathcal{S}_0^r$ be the start set of this snowball search. From there, a variation of backward-snowballing and forward-snowballing[2] steps are applied to construct the set $\mathcal{S}$. That is,

- The set $\mathcal{S}_{-1}$ is obtained by backwards-snowballing on the set $\mathcal{S}_0^r$;
- The set $\mathcal{S}_{-2}$ is obtained by backwards-snowballing on the set $\mathcal{S}_{-1}^r$;
- The set $\mathcal{S}_{+1}$ is obtained by forward-snowballing on the set $\mathcal{S}_0^r$;
- The set $\mathcal{S}_{+1-1}$ is obtained by backwards-snowballing on the set $\mathcal{S}_{+1}^r$;
- The set $\mathcal{S}_{+2}$ is obtained by forward-snowballing on the set $\mathcal{S}_{+1}^r$;
- The set $\mathcal{S}_{+2-1}$ is obtained by backwards-snowballing on the set $\mathcal{S}_{+2}^r$;

and finally the set $\mathcal{S}$ is obtained by taking the union of all the previously mentioned sets, i.e. $\mathcal{S} := \mathcal{S}_0 \cup \mathcal{S}_{-1} \cup \mathcal{S}_{-2} \cup \mathcal{S}_{+1} \cup \mathcal{S}_{+1-1} \cup \mathcal{S}_{+2} \cup \mathcal{S}_{+2-1}$.

Unfortunately, the outlined literature collection process exhibits a rather undesirable property. That is, in an ideal world the methodology would provide a perfectly reproducible algorithm that reliably and deterministically produces the same set of publications on each execution. Unfortunately, this property cannot be satisfied by the constructed methodology, as any employment of forward-snowballing introduces variability into the system. Considering the importance of forward-snowballing to identify the most recent literature, foregoing the application of this tool in the construction of $\mathcal{S}$ would not have been feasible. Therefore, conditions that further infringe upon reproducibility, i.e. the soft categorisation of relevance and the removal of inaccessible publications, seems justifiable.

[2] using Google Scholar between 16.04.2020 and 20.04.2020.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The dataset constructed using this methodology is publicly available.[3] It is stored in a SQLite-database and was constructed using SQLAlchemy. This database can store publications, authors, venues and tags. Its structure is depicted in Figure 2.1 as an ER-diagram using the notation in [Che76].

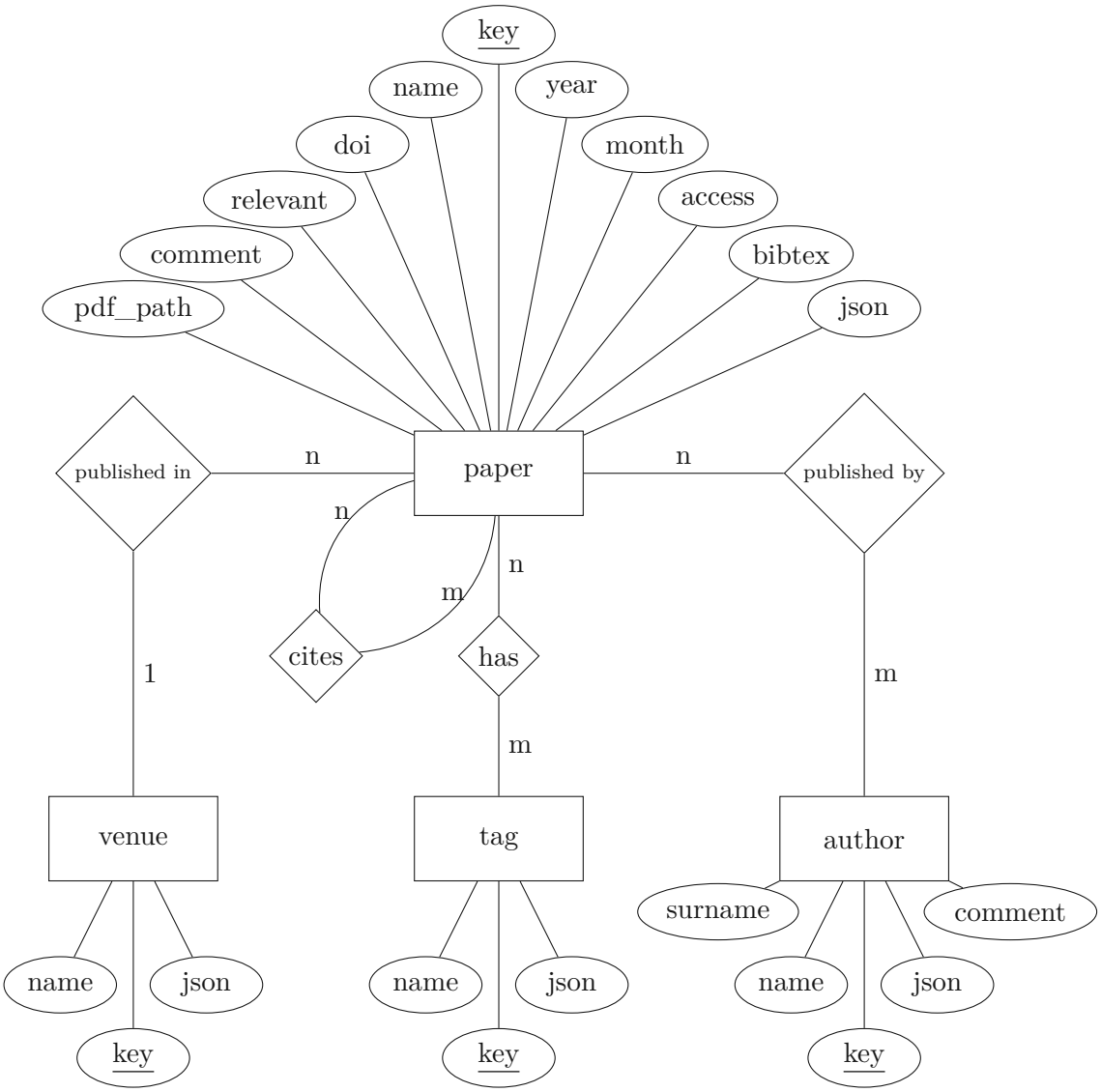


Figure 2.1: ER-Diagram of the database.

The table "Venue" stores platforms that publish research, e.g. journals such as AI or conferences such as IJCAI. This table is rather sparse, having only a column for the names of the venues and a column reserved for a JSON-string in case additional data

[3] https://github.com/KonstantinRK/CausalitySurvey

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

must be stored. Similar in structure is the "Tag" table. Its intended purpose is to store tags that can be used to further categorise publications, e.g. they are used to identify which publications were added to the database at which snowballing step. Slightly more complex is the "Author" table, which contains an additional column for surnames, as well as a column reserved for comments. The last column allows one to differentiate authors with identical names. By far the most extensive one is the Table "Paper". Firstly, it provides columns for basic information such as the title, the DOI, the publication year and month, as well as its BibTeX-string. Moreover, it contains columns to track whether a publication was accessible and whether it is deemed relevant. Furthermore, for easy access there exists also a column storing the local path to the pdf-file of the publication. Lastly, columns for comments and for storing additional data in JSON-format are provided as well.

Those tables relate to each other as follows. Firstly, the table "Venue" is connected to the table "Paper" via an 1:n-relation. Secondly, the "Author"- and the "Paper"-table are in an m:n-relation. The same holds true for the "Tag"- and the "Paper"-table. Thirdly, to store the citation relations between publications, the table "Paper" is in an m:n-relation with itself.

During the data collection, publications will be assigned one of seven tags. Those are `0`, `-1`, `-2`, `+1`, `+1-1` `+2` and `+2-1` and indicate membership to the respective sets constructed during the snowballing process.

2.1.2 Data Analysis Methods

The data analysis was conducted according to the following steps. The first is the construction of several graphs using the stored citation relation and the co-authorship relation. This is followed by a heuristic detection of research communities obtained through the application of a community detection algorithm. Following this, a combination of centrality measures and publication counts will be used to identify the most prominent authors in this field. Lastly, a similar approach will be employed to identify its most significant publications.

Considering the collected data it is possible to derive several graph structures. The first two are graphs have vertices that represent publications and edges that correspond to citations, i.e. each edge (A, B) implies that there exists a publication A that references another publication B. Therefore, they are directed graphs[4]. The first graph $\mathcal{G}_p$, called the "publication graph", is build using the publications in $\mathcal{S}^r$ as its vertices. Hence, it contains only those publications that were deemed relevant and that have been published from 2010 onwards. $\mathcal{G}_p$ is of particular importance because it serves as the bedrock of all subsequent analysis. That is, $\mathcal{G}_p$ is used to construct a set of important publications from which all formal languages, definitions and examples discussed in subsequent chapters are extracted. The second graph $\mathcal{G}_f$, called the "full graph", uses $\mathcal{S}$ as its vertex set,

[4]These graphs contain cycles. This is due to the fact that sometimes not yet publicised works are cited, which contain a reference back to the original publication.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

thus it contains all publications irrespective of their publication date and their relevance marker. The purpose of this graph is to hedge against the limited perspective provided by $\mathcal{G}_p$. That is, the quick analysis of $\mathcal{G}_f$ allows the detection of additional pointers to potentially relevant literature, thus providing additional insight into the literature before 2010, as well as opening up the possibility of identifying relevant publications that were wrongly classified as irrelevant during the snowballing procedure. Unfortunately, by construction only the publications in $\mathcal{S}$ that were published after (and including) 2010 contain a complete mapping of their citations, thus any results obtained from $\mathcal{G}_f$ are inherently of inferior quality, as the incomplete citation relation obviously distorts the relative relevance between the analysed publications. Hence, while sufficient for hedging, this thesis refrains from indulging in any further analysis of $\mathcal{G}_f$.

Three additional graphs are also produced. They encode information about the authors and the relationships between them. Starting with the "author graph" $\mathcal{G}_a$, which encodes the citations between authors, rather than between publications. Every vertex in $\mathcal{G}_a$ represents an author, who has at least one publication represented in $\mathcal{G}_p$. Every edge in $\mathcal{G}_a$ between an author A and an author B indicates that there exists at least one publication of A that cites at least one publication of B (with respect to $\mathcal{G}_p$), while its weight represents the frequency of this occurrence. To account for multiple citations, as well as self-referential behaviour, $\mathcal{G}_a$ must be a directed, weighted graph containing loops. The purpose of this graph is to identify potentially influential figures, whose ideas shape the discourse around the formal approaches to token causality.

The second, $\mathcal{G}_c$ or "collaboration graph", encodes the collaborations between authors. That is, while its vertex set is identical to $\mathcal{G}_a$'s vertex set, an edge in $\mathcal{G}_c$ between an author A and an author B, indicates that A and B co-authored a publication in $\mathcal{G}_p$, thus the weight of an edge represents how often they collaborated on publications in $\mathcal{G}_p$. $\mathcal{G}_c$, is primarily used as an intermediary step in the analysis procedure. However, it will be included in the analysis, as it provides an overview of the collaboration relations among the authors, allowing for the detection of research communities in a strict sense. The last, $\mathcal{G}_m$ or "merged graph" is simply a merger of $\mathcal{G}_a$ and $\mathcal{G}_c$, where each undirected edge is replaced by two opposing directed edges (in the case of duplicate edges, their weights are summed up). This graph will be used to detect research communities within the literature that produce a high relative volume of relevant publications. In the ideal case, this should uncover thematic clusters, aiding the reader in the search for related literature. To summarise, $\mathcal{G}_f$ and $\mathcal{G}_p$ are directed graphs; $\mathcal{G}_a$ and $\mathcal{G}_m$ are directed, weighted graphs containing loops and $\mathcal{G}_c$ is an undirected, weighted graph.

The constructed graphs are subjected to two different kinds of information extraction processes. The first uses the community detection algorithm to identify research communities, while the second uses a variety of centrality measures to identify important publications and authors. Approaches leveraging such techniques fall under the term "citation analysis". An area of research concerned with the discovery and management of literature by analysing its references to evaluate scholarly contributions, track the flow of knowledge, study the structure of the research field, etc. [ZS15, p. 1-5]. While useful, this

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

requires the acceptance of several assumptions. Namely, citation of a document implies use of that document by the citing author; citation of a document (author, journal, etc.) reflects the merit (quality, significance, impact); citations are made to the best possible works; a cited document is related in content to the citing document; all citations are equal.

While accepting those rather strong assumptions is problematic, many of which were already violated by this text, additional concerns with this technique arise when one also considers that there can be various problems in the data, e.g. errors, self-citations or multiple authors. However, due to the fact that those techniques are only used in a rudimentary manner to provide a starting point for the subsequent research, a proper justification of the applicability of those assumptions with respect to the given data will be omitted. See [Smi81] for a detailed discussion about the validity of those assumptions.

Firstly, the detection of communities. For the purposes of this work, a group of researchers is classified as a community (with respect to their work on causation), if the group is of size greater than two and if the group produced more than two relevant publications. This approach should provide a rough estimate of the research clusters in the literature based on the information encoded in the merged graph. The community detection itself is accomplished using an algorithm published in [RB08], which is suitable for any directed, weighted graphs. Hence, if one excludes self-referential behaviour, the same algorithm can be used on any graph up for analysis, thus it is well suited for $\mathcal{G}_m$. Furthermore, [RB08] introduces their algorithm by studying a citation graph, thus demonstrating the suitability of the algorithm for such tasks. See [RB08] for further discussion and additional details. Nevertheless, to identify relevant communities the following procedure will be used. Firstly, $\mathcal{G}_m$ will be cleared from all loops. Secondly, all authors that are only cited or that only cite, with respect to the collected data, are removed. Thirdly, the community detection algorithm is applied to the graph. Lastly, all communities are ranked based on the average number of relevant publications per author. Hence, providing the possibility for smaller communities to get some spotlight as well.

Secondly, the primary technique used in this work to assess the importance of a vertex in a graph relies on the use of centrality measures. Being significant for the work in subsequent chapters, the discussion of such measures warrens a more detailed discussion as compared to community detection part. In general, those centrality measures are used to rank vertices based on some notion of importance. In particular, they can be used to understand diffusion processes, assess an individuals risk of infection or explain the influence of a person in a social network [BJT19]. According to [dPMGAO11] degree, closeness and betweenness centralities are the most popular ones. Additionally, there exists a family of centralities that is closely tied to the field of spectral graph theory (see [Spi12]), as those centralities use eigenvalues and eigenvectors in their computation. This includes measures such as the eigenvector centrality, alpha centrality, page rank and Katz-Bonacich centrality (see [BJT19]).

The following provides a brief intuition about some of the mentioned centrality measures and is compiled from information found in [SR15, dPMGAO11, BJT19, BL01, PBMW99].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The degree centrality is a local measure of importance based on the degree of a vertex. In the case of a weighted graph, the weighted degree of a vertex is taken for this measurement, thus implying that the weight of an edge must reflect some notion of similarity. That is, a higher weight implies a stronger connection, e.g. number of interactions. If one is confronted with a directed graph, the degree centrality dissolves into an in- and out-degree centrality. Although it can be used to assess the "popularity" of a vertex, due to its locality it neglects the remaining structure of the graph. The closeness centrality is computed using the sum of all shortest path lengths. Hence, it is a measure for assessing the importance of a vertex based on how quickly such a vertex can reach every other vertex. Moreover, this implies that edge weights must represent dissimilarity, e.g. distance between vertices. Unfortunately, this measure requires the graph to be strongly connected. Hence, it is not suitable for directed graphs in general. The betweenness centrality gives higher values to vertices that are part of many shortest paths between pairs of vertices. Meaning it attempts to assess the importance of a vertex based on how vital a vertex is for the flow of information between the other vertices in the graph. As this centrality builds upon the notion of the shortest path, it requires the weights of a graph to represent a dissimilarity between vertices. However, a benefit of this centrality is that it is suitable for both directed and undirected graphs. The eigenvector centrality is similar to the degree centrality, as it assesses the importance of a vertex based on the number of neighbours, thus it requires weights to denote similarities. However, it differs in the evaluation of those neighbours, determining the importance of a vertex based on the importance of the vertices in its neighbourhood. That is, the eigenvector centrality is computed by assuming that the centrality of a vertex is proportional to the sum of eigenvector centralities of the vertex's neighbours. Hence, it is a self-referential process. Unfortunately, common implementations require graphs to be undirected and connected. The alpha centrality is a generalisation of the Eigenvector centrality for directed graphs. The idea behind this centrality is that it assumes that a vertex has some exogenously defined start value. The Katz-Bonacich centrality generalises the Eigenvector centrality by reducing the importance of distant vertices. Page Rank relativises the centrality score passed on by a vertex, based on the number if neighbours. Hence, a vertex having a directed edge to an important vertex must not necessarily have high importance itself, e.g. a webpage linking to an important webpage must not necessarily be important itself. Furthermore, to ensure sensible results in directed graphs, dead ends are avoided by jumping to a random vertex instead.

Each of those measures allow for a separate ranking of publications and authors. However, a blind application of those methods would neglect the structure of the graphs and thus could lead to erroneous results. Hence, some additional care must be given and some slight adjustments to the graphs are required. In the case of $\mathcal{G}_p$, one is faced with a directed (and not strongly connected) graph. Therefore, the closeness or eigenvector centrality are ill-suited for application on this graph. Furthermore, the degree centrality is obviously applicable. However, it decomposes into two separate measures. Additionally, the regular degree centrality will be used as well. That is, in the context of this particular dataset, the regular degree distribution actually provides a rough compromise between

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the recency bias of the out-degree, as well as the conservative tendencies associated with the in-degree measure (this can be observed in Figure 6.1). Hence, the undirected degree measure will be used as well. Unfortunately, due to their locality degree centralities provide a rather limiting picture, thus in an attempt to compensate for this shortcoming an alternative to the eigenvector centrality, namely page rank, is used. Although it is somewhat unusual to use this algorithm for citation graphs, this approach is not unheard of, see [DYFC09, MGZ08, CXMR07, MR08, NJFD14] for an in-depth discussions. One particular benefit of determining the importance of a publication in such a manner is that under Page Rank simply referencing an important publication, does not indicate a publications own importance. The last remaining common centrality measure, the Betweenness centrality, can and thus will be applied. Providing yet another dimension for selecting publications.

In the case of the author graph $\mathcal{G}_a$, one is faced with a directed, weighted graph containing loops. Hence, modifications to the graph are required. Firstly, while included for the sake of completion in the graph $\mathcal{G}_a$, it seems sensible to discount self-referential behaviour for the ranking. Secondly, all centrality measures used in the ranking of publications can accommodate weights in their assessment. However, the Betweenness Centrality requires the weight of an edge to express dissimilarity, thus it is necessary to convert the weights such that they express dissimilarity rather than similarity between vertices (see [Run12, p. 13]). Moreover, in addition to the centralities it is reasonable to include the number of publications (in the examined field) as an additional measure.

2.2 Analysis

This section presents the results obtained by instantiating the methodology outlined in Section 2.1. This includes a documentation of the publication collection process, a rudimentary presentation of the constructed graphs, as well as their analysis utilising the described methods to generate pointers to possibly relevant publications, authors and communities. Moreover, the most important aspect of this section is the construction of a set of seemingly important publications, whose content are surveyed and discussed in subsequent chapters. Hence, Section 2.2.1 contains the documentation of the publication collection process, presents the constructed graphs and describes the collected literature through the performance of a rudimentary quantitative analysis. Furthermore, Section 2.2.2 executes the analysis of the constructed graphs.

2.2.1 Data Preparation and Basic Analysis

Here the data collection process is quantitatively described. Additionally, any detected mistakes as well as deviations from the constructed methodology are highlighted as well. Moreover, this is followed by a brief and offensively basic investigation into the collected data and the constructed graphs.

Using the outlined methodology the following literature database was constructed. By

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

collecting all publications from the venues Journal Artificial Intelligence (AI), Journal Artificial Intelligence and Law (AI&Law), International Joint Conferences on Artificial Intelligence Organization (IJCAI) and Knowledge-Bases Systems (KBS) that were published between 01.2017 and 3.2020 one obtains a set containing 4223 publications. To be precise, AI contributed 267 publications, by contrast, AI&Law provided only 60. Furthermore, from KBS a total of 1281 publications could be obtained, while the majority of publications, i.e. 2615, was sourced from IJCAI. After applying the keyword based filter one obtains $\mathcal{S}_0$, which contains 37 publications only.[5]

After closer investigation, the publications deemed relevant according to the specified criteria are

- Proof with and without probabilities [Ver17];
- Characterizing causal action theories and their implementations in answer set programming[6]; [ZL17];
- Actual Causality in a Logical Setting [Boc18a];
- On the conditional logic of simulation models [II18];
- Counterfactual Resimulation for Causal Analysis of Rule-Based Models [LYF18];
- Scalable Probabilistic Causal Structure Discovery [SPG18];
- ASP-based discovery of semi-Markovian causal models under weaker assumptions [ZZE+19];
- Arguing about causes in law: a semi-formal framework for causal arguments [LSW19b].

Hence, $\mathcal{S}_0^c$ contains only 8 publications. Executing the described snowballing steps on the start set, one obtains a total of 872 publications. Out of which only 294 (around 34%) are categorised as relevant. As depicted in Figure 2.2, this can be made more precise. Meaning that $\mathcal{S}_{-1}$ obtained by performing backward-snowballing on $\mathcal{S}_0^r$, contains 204 publications out of which only 79 are relevant. The second backward-snowballing step, provided a total of 486 publications with 165 being relevant. The set $\mathcal{S}_{+1}$ contains 30 publications collected by forward-snowballing on $\mathcal{S}_0^r$. Performing a backward-snowballing step on $\mathcal{S}_{+1}^r$ generates $\mathcal{S}_{+1-1}$ which contains 63 publications from which 25 are deemed relevant. The second forward-snowballing step, produces $\mathcal{S}_{+2}$ resulting in an additional

[5]Those publications are [vdZLT19, Ver17, Che19, NFLG19, LFWZ19, LWFZ18, ZLW+18, CF17, LWZ17, ZL17, Mu18, KOP19, HSJ17, ZB17, ZHZ+17, LJE+17, SOM17, ZWW16, AR16, CGS+18, Boc18a, II18, LYF18, CF18, ZWW18, BJO18, JZB18, SPG18, WLC18, XWY+19, ZZE+19, CQZ+19, SG19, XM19, HBF+19, SSS+19, LSW19b]

[6]Since [ZL17] was not accessible "Characterizing causal action theories and their implementations in answer set programming: Action languages b, c, and beyond" [ZL15] will be used for the snowballing step. This departure from the methodology, is justified due to its initially high reprieved relevancy.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

7 publications with only 3 relevant ones among them. Lastly, by performing a final backward-snowballing step on $\mathcal{S}^r_{+2}$, 45 new publications are discovered increasing the number of relevant publications by another 7.

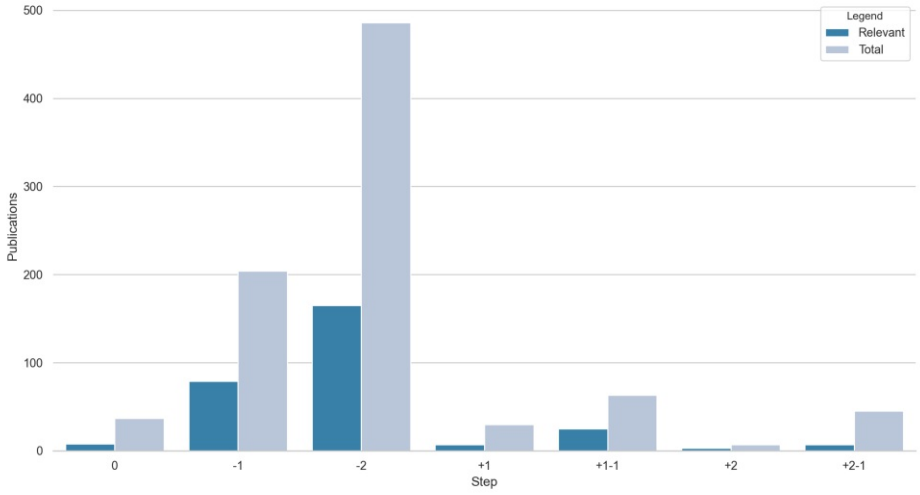


Figure 2.2: Number of publications and relevant publications added during each snowballing step. (From left to right $\mathcal{S}^c_x/\mathcal{S}_x$: 8/37, 79/204, 165/486, 7/30, 25/63, 3/7, 7/45)

Although great effort was taken to make the data collection sufficiently accurate. Some mistakes were discovered after the data collection was already completed. Particularly of note is that there are two publications titled "Causes and explanations: a structural-model approach: part i: causes" by the same authors, i.e. [HP01] and [HP05]. During the data collection process, those two publication were unfortunately conflated.

From the meta-data of the publications alone, one can observe the contributions to this field over the years. That is, given the publication dates of the literature collected in $\mathcal{S}^c$ it is possible to construct Figure 2.3, which depicts the distribution of the number of publications (in $\mathcal{S}^c$) per year across the past 50 years. Furthermore, according to the data collected, the decade between 2000 and 2010 was the most productive period, i.e. $\mathcal{S}$ contains 73 publications before 2000, 114 between 2000 and 2010 and 107 publications from 2010 onwards. Additionally, it can be observed that 2004, 2007 and 2009 were the most productive years overall. Containing notable publications such as "Nonmonotonic Causal Theories" [GLL+04], "Causes and Norms" [HK09], "Prevention, Preemption, and the Principle of Sufficient Reason" [Hit07a], "Two Concepts of Causation" [Hal04] and "Structural Equations and Causation" [Hal07] [7].

[7] discussing and using formalisms such as Neuron Diagrams, Structural Equations and some variant of McCain and Turner's Causal Logic

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

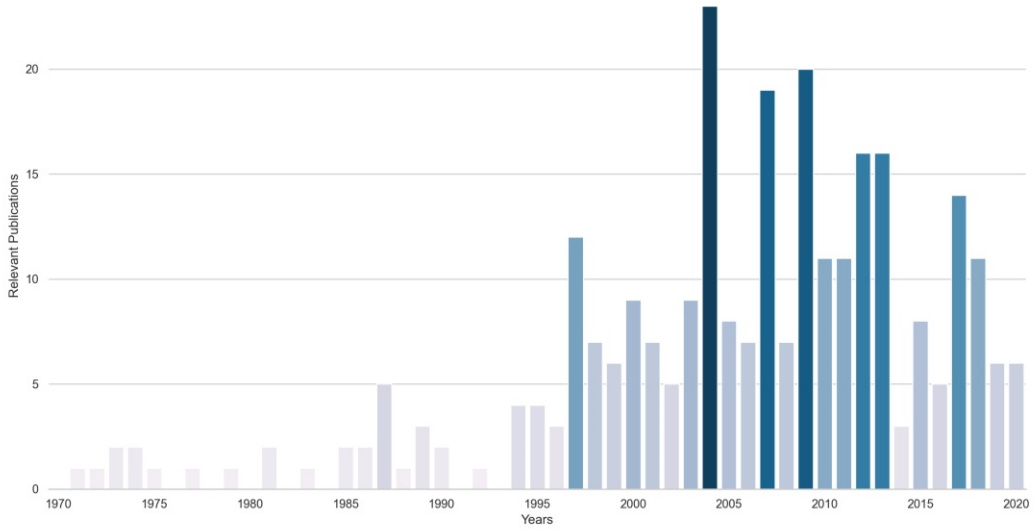


Figure 2.3: Number of relevant publications per year (a negligible amount publications occur before 1970)

Further information can be extracted by encoding the collected data as graphs. As discussed earlier, the set of publications $\mathcal{S}$ and their references naturally induce a directed graph containing 872 vertices and 2052 edges. This graph, i.e. $\mathcal{G}_f$ can be observed in Figure 2.4. Induced by the set $\mathcal{S}^r$, containing publications that are both relevant and are published after (and including) 2010, one obtains $\mathcal{G}_p$ as a sub-graph of $\mathcal{G}_f$. $\mathcal{G}_p$ contains only 107 and 326 edges and can be observed in Figure 2.5. Using $\mathcal{G}_p$ one can then compute $\mathcal{G}_a$, visible in Figure 2.6, which contains a total of 130 vertices and 462 edges. As discussed this graph encodes the citations between authors and not the one between publications. Hence, it requires that its directed edges are weighted. To analyse the co-authorship relation one can create $\mathcal{G}_c$, which is depicted in Figure 2.7 and contains 130 vertices and 192 undirected, weighted edges. Lastly, $\mathcal{G}_m$, depicted in Figure 2.8, is the merger of $\mathcal{G}_a$ and $\mathcal{G}_c$, thus it contains 130 vertices and 755 edges. For a quick overview of some of their basic properties please consult Table 2.1 and Table 2.2, as well as Figure 2.9, Figure 2.10 and Figure 2.12 respectively.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

	Vertices	Edges	Density	Clustering Coefficient
$\mathcal{G}_p$	107	326	0.0287	0.2791
$\mathcal{G}_a$	130	462	0.0275	0.3144
$\mathcal{G}_c$	130	192	0.0229	0.7843
$\mathcal{G}_a$	130	755	0.045	0.5141

Table 2.1: General properties of the discussed graphs. Other common measures such as average path length, radius and diameter, as well as vertex- and edge connectivity are omitted as all graphs in question are disconnected.

	Minimum	Maximum	Average	Median
$\mathcal{G}_p$				
Degree	1	22	6.09346	5
In-Degree	0	18	3.04673	2
Out-Degree	0	13	3.04673	2
$\mathcal{G}_a$				
Degree	0	50	7.10769	4
In-Degree	0	29	3.55385	3
Out-Degree	0	23	3.55385	0
Weighted Degree	0	85	10.2615	5
Weighted In-Degree	0	53	5.13077	3
Weighted Out-Degree	0	32	5.13077	0
$\mathcal{G}_c$				
Degree	0	13	2.95385	2
Weighted Degree	0	17	3.67692	2.5
$\mathcal{G}_m$				
Degree	0	53	11.6154	10
In-Degree	0	29	5.80769	5
Out-Degree	0	24	5.80769	4
Weighted Degree	0	97	17.6154	13
Weighted In-Degree	0	59	8.80769	5.5
Weighted Out-Degree	0	32	8.80769	5.5

Table 2.2: Degree Statistic of $\mathcal{G}_p$, $\mathcal{G}_a$ and $\mathcal{G}_c$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

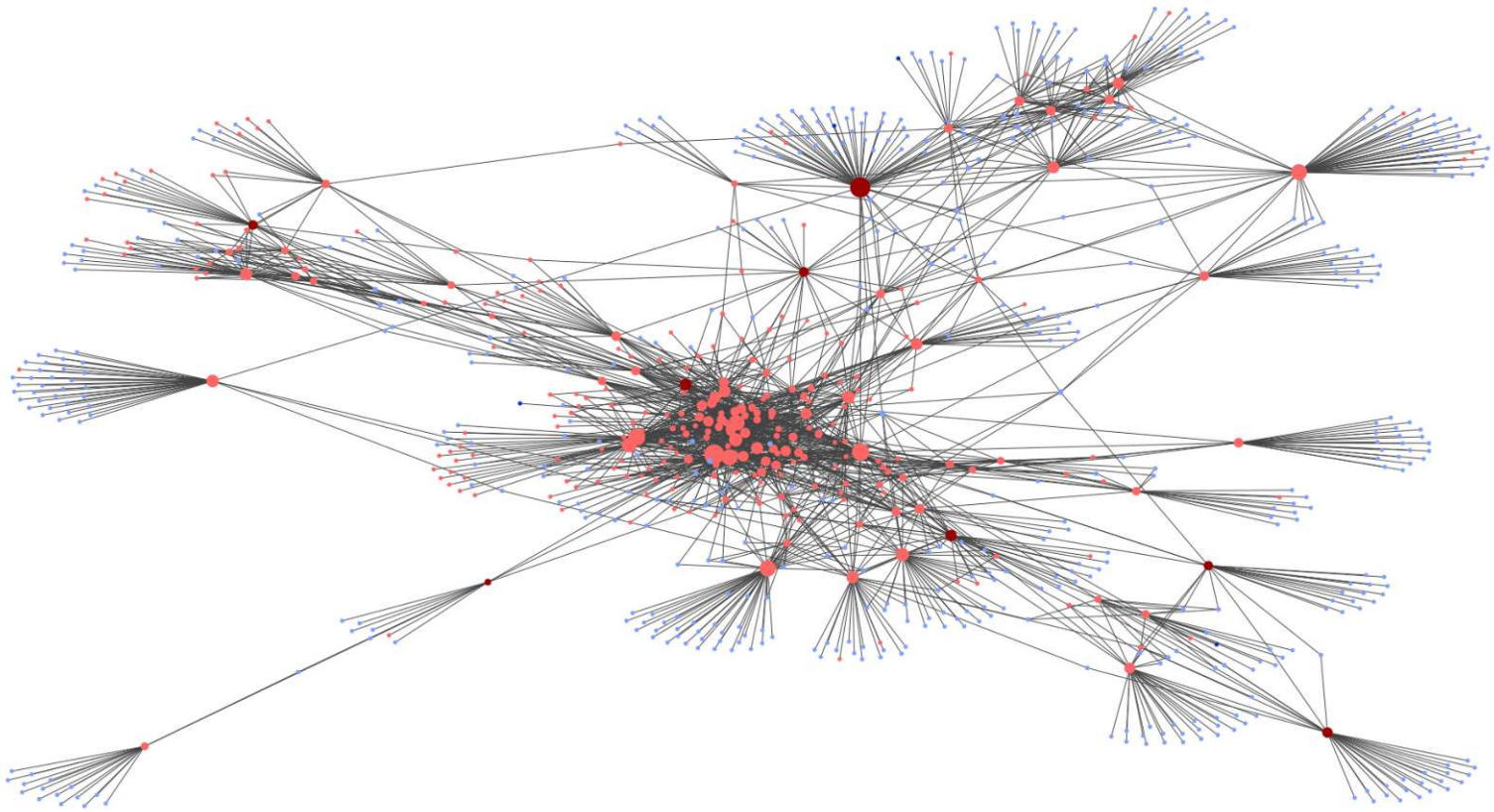


Figure 2.4: $\mathcal{G}_f$ where dark red (dark blue) indicates a(n) (ir)relevant publication in $\mathcal{S}_0$ and light red (light blue) indicates a(n) (ir)relevant publication in $\mathcal{S} \setminus \mathcal{S}_0$. (Isolated vertices are not depicted in this graph and edge direction is suppressed.)

Figure 2.5: The publication graph $\mathcal{G}_p$, containing all relevant publications published after (and including) 2010, which consists out of 107 vertices (8 are in $\mathcal{S}^r_0$; 29 are in $\mathcal{S}^r_{-1}$; 42 are in $\mathcal{S}^r_{-2}$; 7 are in $\mathcal{S}^r_{+1}$; 17 are in $\mathcal{S}^r_{+2}$; 3 are in $\mathcal{S}^r_{+1-1}$; 1 are in $\mathcal{S}^r_{+2-1}$) and 326 edges. The vertex size correlates with vertex degree. 1: Graded Causation and Defaults; 2: Actual Causation: A Stone Soup Essay; 3: Cause without Default; 4: Actual Causation and the Art of Modeling; 5: Actual Causality in a Logical Setting; 6: Counterfactuals; 7: A Partial Theory of Actual Causation; 8: From Programs to Causal Models; 9: A Modification of the Halpern-Pearl Definition of Causality; 10: Explaining Actual Causation in Terms of Possible Causal Processes; (Isolated vertices are not depicted in this graph and edge direction is suppressed.)

TU Bibliothek Your knowledge hub WIEN
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

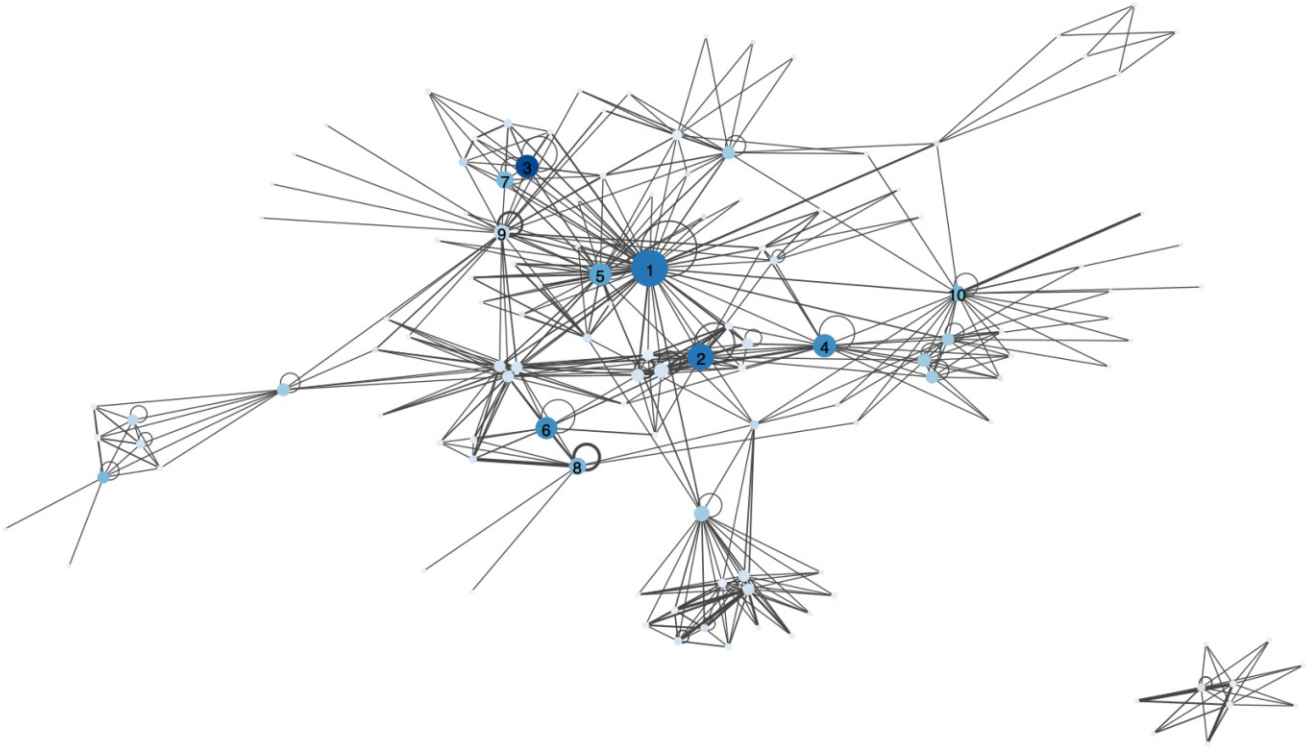


Figure 2.6: The author graph $\mathcal{G}_a$ based on $\mathcal{G}_p$, which consists of 130 vertices and 462 edges. Darker colors indicate a higher number of publications in $\mathcal{G}_p$. Vertex size correlates with the weighted vertex degree; edge width correlates with edge weight. 1: Halpern; 2: Lagnado; 3: Vennekens; 4: Gerstenberg; 5: Hitchcock; 6: Bex; 7: Beckers; 8: Verheij; 9: Bochman; 10: Icard (Isolated vertices are not depicted in this graph.)

Figure 2.7: The collaboration graph $\mathcal{G}_c$, based on $\mathcal{G}_p$, which consists of 130 vertices and 192 edges. Darker colors indicate a higher number of publications in $\mathcal{G}_p$. Vertex size correlates with the weighted vertex degree. Edge width correlates with edge weight. An edge with color red has weight greater than 1. (1: Lagnado; 2: Eberhardt; 3: Gerstenberg; 4: Vennekens; 5: Zhang; 6: Goodman; 7: Tenenbaum; 8: Fontana; 9: Lee; 10: Lifschitz.)

TU Bibliothek Your knowledge hub WIEN
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

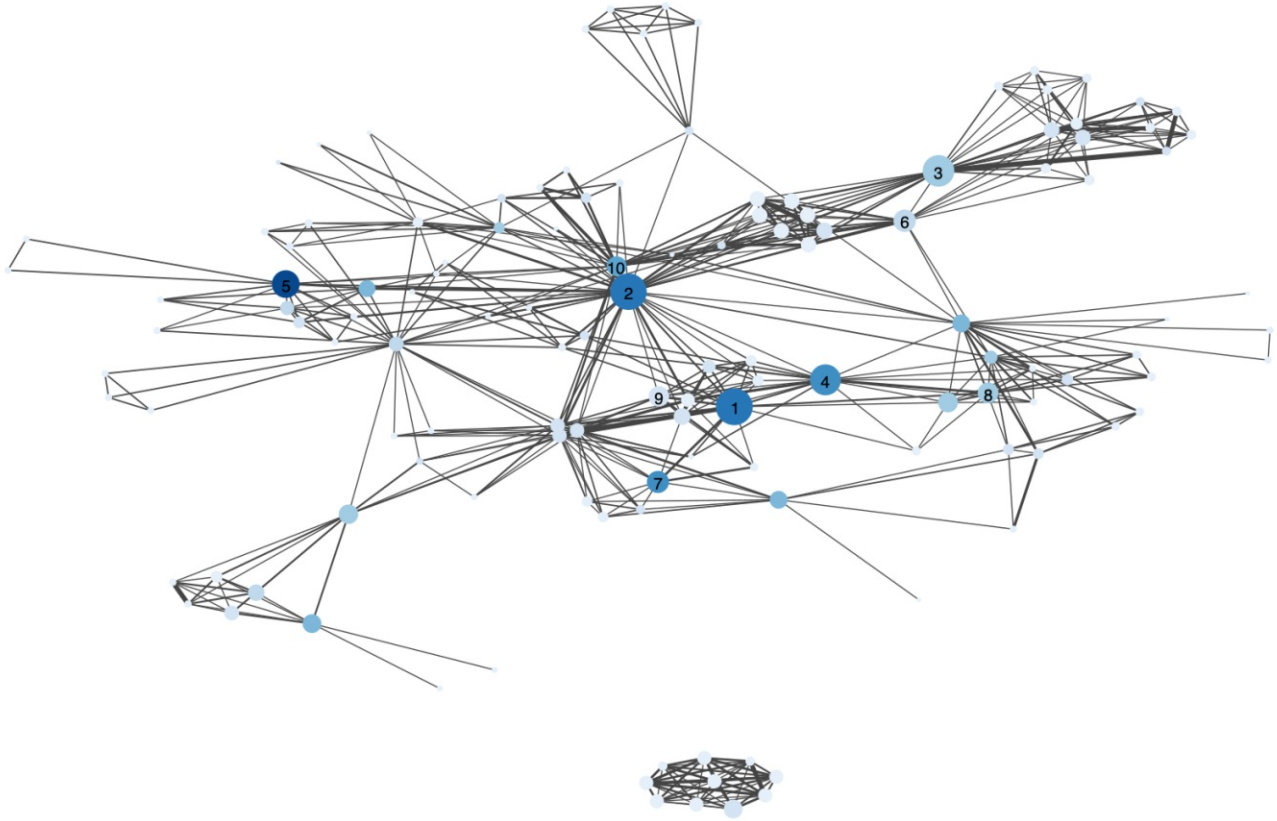


Figure 2.8: The merged graph $\mathcal{G}_m$ based on $\mathcal{G}_p$, which consists of 130 vertices and 755 edges. Darker colors indicate a higher number of publications in $\mathcal{G}_p$. Vertex size correlates with the weighted vertex degree; edge width correlates with edge weight. 1: Lagnado; 2: Halpern; 3: Eberhardt; 4: Gerstenberg; 5: Vennekens; 6: Zhang; 7: Bex; 8: Tenenbaum; 9: Chockler; 10: Hitchcock. (Isolated vertices are not depicted in this graph and edge direction is suppressed.)

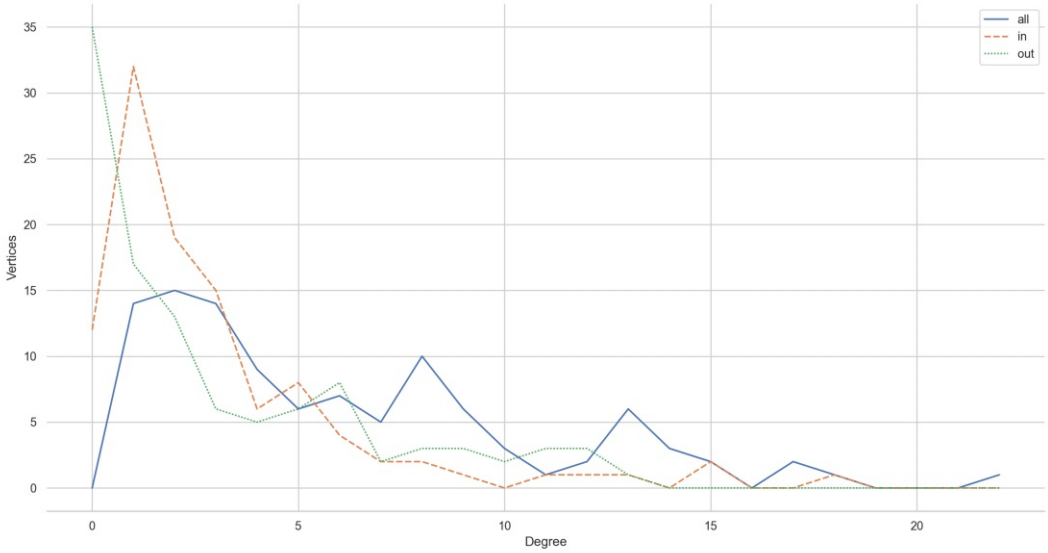


Figure 2.9: A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_p$

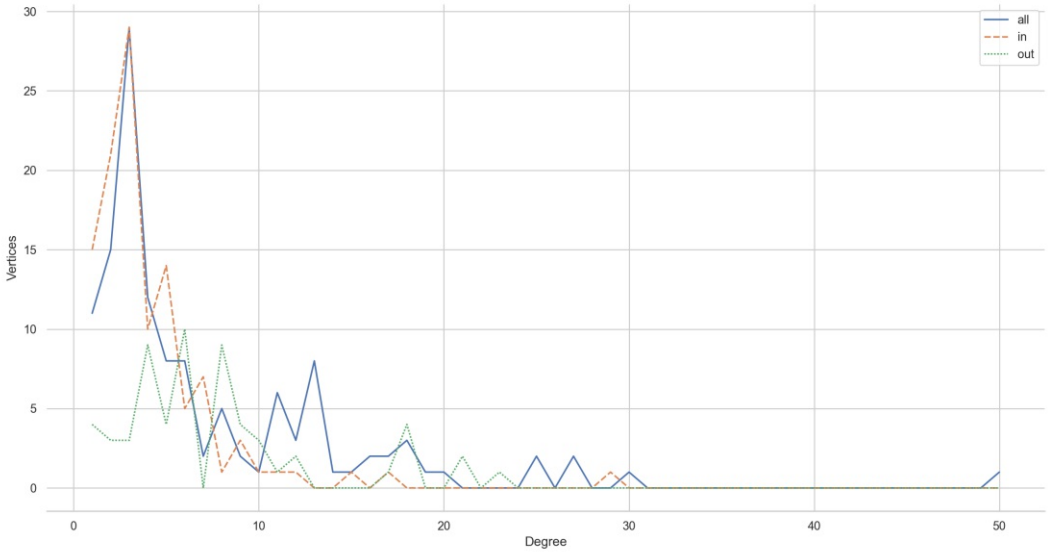


Figure 2.10: A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_a$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

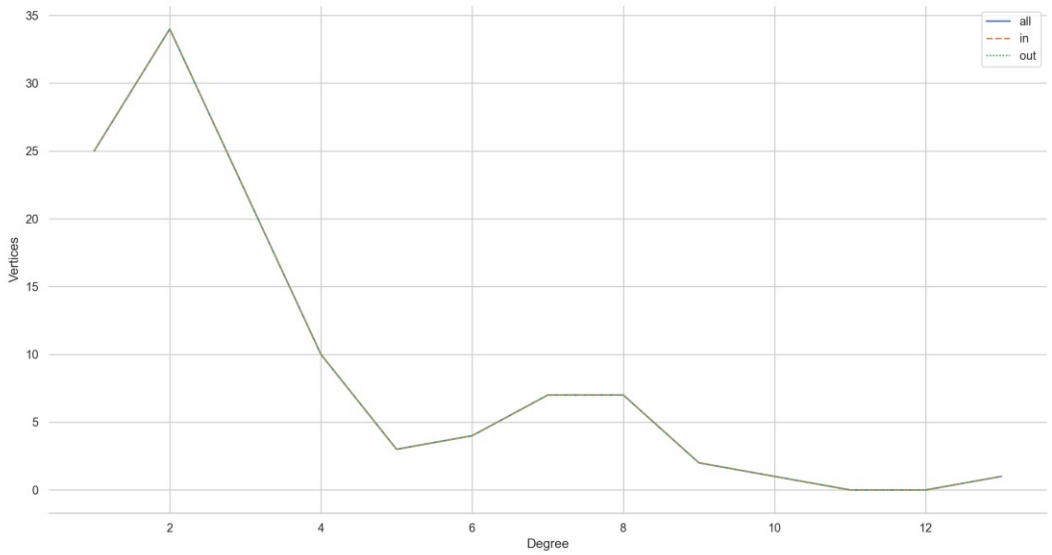


Figure 2.11: A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_c$

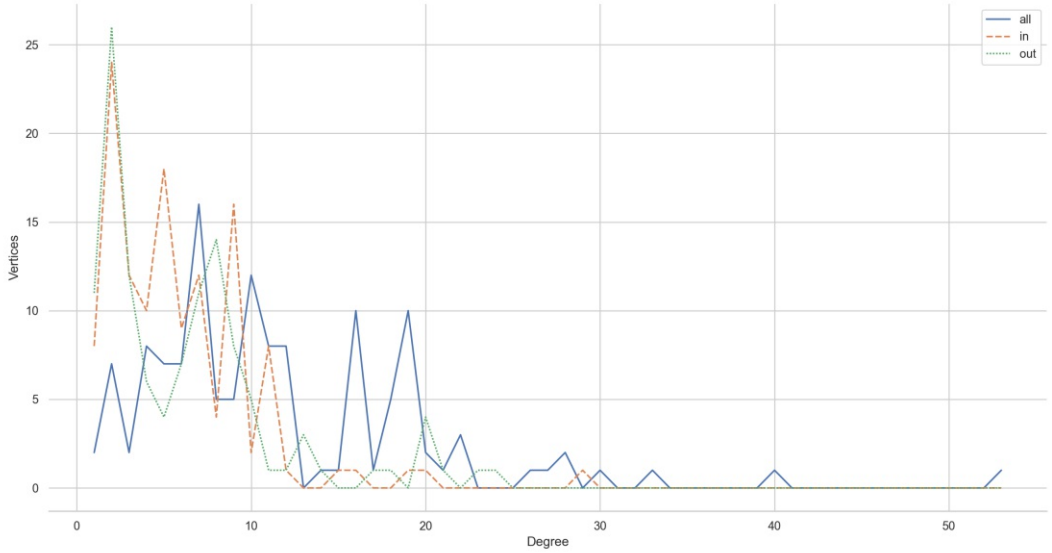


Figure 2.12: A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_m$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2.2.2 Communities, Authors and Publications

Here the results obtained by using the tools introduced in Section 2.1.2 to analyse the data discussed in Section 2.2.1, which were collected according to the methodology in Section 2.1.1 are presented. The discussion starts by presenting the results of the community detection algorithm, is followed by providing a selection of authors deemed important and concludes by building the set of important publications whose content will be described and categorised in the subsequent chapters.

Applying [RB08]'s community detection algorithm to a subgraph of $\mathcal{G}_m$, which was obtained by removing all vertices with zero in- or zero out-degree, the following communities could be identified. Recall that a grouping of researchers is classified as a community only if the group has a size greater than two and if the sum of their relevant publications exceeds two. As listed in Table 2.3 , the algorithm detected eight communities. As a whole, they can be viewed in Figure 2.13 while the connections within each individual group can be viewed in Figure 6.2-6.9.

When ranked based on the number of relevant publications per author, there are three groups, namely Group 4, 7 and 8, that set themselves apart from the remaining groups by having disproportionally high relevancy. Starting with the lowest of those three, Group 8. This group consists of 4 people, Ibeling, Icard, Kominsky and Knobe, with Icard the most prominent author of this group. They published 6 relevant publications, most of those publications discuss to some extent or another a language called simulation models, which can be used to encode causal relationships. Additionally, they discuss other formal languages such as causal models and Bayesian networks. The second research community is Group 7 and consists of five people, i.e. Verheij, Bex, Walton, van Koppen and Prakken. They contributed a total of eight relevant publications, all of which are placed within the context of causality and law. Furthermore, applications of Bayesian networks and a heavy emphasis on stories can be detected. The first one, Group 4, consisting of a total of 16 people who together are responsible for 31 relevant publications. One unifying aspect exhibited by many of the publications in this community, is the emphasis on logic, as well as their attempts to formalise token causality from an inductive example first approach. However, being of considerable size this group can thematically be further segmented. In particular, one cluster seems to emerge around Vennekens, discussing a variety of approaches to causation with the most notable ones being based on a formal language called CP-Logic. Another is thematically grouped around causal models, where Halpern seems to be the most dominant influence. While those are the two main areas discussed in Group 4, this is by no means exhaustive. For example, the work of Bochman, while being subject wise in closer proximity to the community around Halpern, cannot be placed in either of the two groups with absolute certainty. A more significant failure of the employed heuristic can be observed with the work of Cabalar and Fandinno, which is thematically closer to the research conducted by Group 5, which investigates causality in the context of logic programming.

To identify the most important authors, six different rankings are established. That is,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Nr.	Size	Relevancy	Relevant Publications	Authors
1	15	0.53	8	Zhang Jiji, Eberhardt Frederick, Mayer Wolfgang, Li Mark Junjie, Baumgartner Michael, Hyttinen Antti, Hoyer Patrik O, Jarvisalo Matti, Glymour Clark, Danks David, Glymour Bruce, Ramsey Joseph, Scheines Richard, Spirtes Peter, Teng Choh Man
2	15	0.8	12	Goodman Noah D, Tenenbaum Joshua B, Gerstenberg Tobias, Chockler Hana, Fenton Norman, Keppens Jeroen, Lagnado David A, Neil Martin, Tenenbaum Josh, Ullman Tomer D, Aleksandrowicz Gadi, Ivrii Alexander, Zultan Ro'i, Lake Brenden M, Gershman Samuel J
3	5	0.6	3	Livengood Jonathan, Alicke Mark D, Rose David, Bloom Dori, Sytsma Justin
4	16	1.94	31	Bochman Alexander, Beckers Sander, Vennekens Joost, Blanchard Thomas, Schaffer Jonathan, Halpern Joseph Y, Hitchcock Christopher, Bruynooghe Maurice, Denecker Marc, Weslake Brad, Huber Franz, Bogaerts Bart, Cabalar Pedro, Fandinno Jorge, Leblanc Emily, Balduccini Marcello
5	9	0.89	8	Zhang Haodi, Lin Fangzhen, Ferraris Paolo, Lee Joohyung, Lierler Yuliya, Lifschitz Vladimir, Yang Fangkai, Casolary Michael, Bartholomew Michael
6	6	0.5	3	Santorio Paolo, Romoli Jacopo, Wittenberg Eva, Ciardelli Ivano, Zhang Linmin, Champollion Lucas
7	5	1.6	8	Verheij Bart, Bex Floris, Walton Douglas, van Koppen Peter J, Prakken Henry
8	4	1.5	6	ibeling duligur, icard thomas, kominsky jonathan f, knobe joshua

Table 2.3: Communities Overview

three rankings rely on the weighted degree of a vertex, with the second and third being established using the weighted in- and out-degree respectively. The fourth, ranks the vertices according to the results provided by the betweenness centrality, the fifth relies on the values computed by the page rank algorithm and the last measures the importance of an author using their publication count. By aggregating the top 15 authors, see Table 2.4, across all 6 rankings into a single set, 33 important authors can be identified. Among those authors such as Lifschitz, Icard, Bochman, Eberhardt, Hitchcock, Gerstenberg, Lagnado and Halpern consistently score high across each ranking and are thus of particularly of note. Using $\mathcal{G}_c$ one can observe that there have been collaborations between Bochman and Lifschitz; Halpern and Hitchcock; Gerstenberg and Icard. Those collaborations are also reflected in the kinds of approaches those authors ascribe to when studying causation. That is, Both Lifschitz and Bochman focus on variants of the causal theory put forward in [MT$^+$97] and tend to approach causality from a regularity theoretic point of view. By

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

contrast, Halpern and Hichcock strongly adhere to the structural equation framework. Their investigations into causality, while emerging from the counterfactual tradition, recently incorporate some regularity theoretic tools, e.g. extending causal models with normality rankings. As opposed to all other authors mentioned, who tend to approach causality from a more theoretical angle, Gerstenberg and Icard, set themselves apart, by conducting empirical studies investigating how humans form their causal judgements and what role the attribution of responsibility play in those judgements. Additionally, they investigate the role of causation in the legal domain. Moreover, similar to Halpern they tend to follow the counterfactual approach to causation and sometimes use structural equations for their modelling. Eberhardt, who cooperated with Clark Glymour on [GDG+10], focuses on the discovery of causal structures. This includes formalisms such as causal Bayesian networks and seems to align closer with the part of the literature centring around machine learning. Lastly, Icard seems to argue for the need for an expressive formalism that emphasises the for him apparent procedural character of causation, thus his latest publications introduce and discuss simulations models, which share a close relationship to Turing machines. Having the highest number of relevant publications, an honourable mention must be given to Vennekens Joost, who worked with structural equations, CP-logic and action languages.

Moving on to the identification of publications deemed important by the outlined methodology using the data encoded in $\mathcal{G}_p$. For each of the five proposed orderings (i.e. degree centrality, in-degree centrality, out-degree centrality, betweenness centrality, page rank), the 15 publications deemed most important, see Table 2.5, are selected and aggregated into a single set, resulting in a total of 36 unique publications.

Particularly notable are the articles [Wes15], [BS17], [HH11], [GDG+10] and [HH15], all of which are ranked highly across all measures. All publications use causal models as their preferred method of encoding causal relations. While all of those publications take on a counterfactual perspective, [HH11], [Wes15] and [HH15] expand the causal model framework to incorporating some aspects of the regularity theory of the causal model approach. For example, this is accomplished by extending causal models such that they allow for the expression of normality. By contrast, [BS17] argues that being more conservative when selecting appropriate causal models, should be preferred over incorporating additional widgets into the structure of causal models itself. As allowing for defaults in causal models does not only increases their complexity, but also provides too much flexibility. [GDG+10] does not engage with the debate about normality in causal inference. Instead, it heavily criticises the attempt of inductively defining causal models from small examples alone, a strategy employed throughout most of the literature.

Moreover, this set of publications will be increased by adding all publication from $\mathcal{S}_0^r$ and by including all publications with 0 in- or out-degree (wrt. to $\mathcal{G}_p$) that have a higher than average (w.r.t. their respective cohort) degree. This result in a set of 44 publications. Meaning that the publications “Causal Reasoning in a Logic with Possible Causal Process Semantics”, “On the Conditional Logic of Simulation Models”, “Evaluation

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

	Degree Centrality	In-Degree Centrality	Out-Degree Centrality	Betweenness Centrality	Page Rank	Publications
1	Halpern Joseph Y	Halpern Joseph Y	Icard Thomas	Halpern Joseph Y	Claassen Tom	Vennekens Joost
2	Hitchcock Christopher	Lagnado David A	Bochman Alexander	Lagnado David A	Heskes Tom	Halpern Joseph Y
3	Bochman Alexander	Gerstenberg Tobias	Halpern Joseph Y	Eberhardt Frederick	Halpern Joseph Y	Lagnado David A
4	Lagnado David A	Hitchcock Christopher	Liepina Ruta	Gerstenberg Tobias	Lagnado David A	Bex Floris
5	Icard Thomas	Lifschitz Vladimir	Sartor Giovanni	Bochman Alexander	Gerstenberg Tobias	Gerstenberg Tobias
6	Gerstenberg Tobias	Eberhardt Frederick	Wyner Adam	Hitchcock Christopher	Lee Joohyung	Verheij Bart
7	Eberhardt Frederick	Claassen Tom	Hitchcock Christopher	Bex Floris	Lifschitz Vladimir	Icard Thomas
8	Liepina Ruta	Heskes Tom	Ibeling Duligur	Zhang Jiji	Lierler Yuliya	Lee Joohyung
9	Sartor Giovanni	Zultan Ro'i	Blanchard Thomas	Fenton Norman	Yang Fangkai	Beckers Sander
10	Wyner Adam	Hyttinen Antti	Baumgartner Michael	Schaffer Jonathan	Zultan Ro'i	Hitchcock Christopher
11	Ibeling Duligur	Hoyer Patrik O	Schaffer Jonathan	Lifschitz Vladimir	Hitchcock Christopher	Ibeling Duligur
12	Schaffer Jonathan	Jarvisalo Matti	Eberhardt Frederick	Lee Joohyung	Eberhardt Frederick	Eberhardt Frederick
13	Fenton Norman	Fenton Norman	Gerstenberg Tobias	Verheij Bart	Hyttinen Antti	Lifschitz Vladimir
14	Lifschitz Vladimir	Lee Joohyung	Keppens Jeroen	Icard Thomas	Hoyer Patrik O	Schaffer Jonathan
15	Chockler Hana	Lierler Yuliya	Lagnado David A	Blanchard Thomas	Jarvisalo Matti	Goodman Noah D

Table 2.4: Top 15 authors according to the Degree Centrality, the In- and Out-Degree Centrality, the Betweenness Centrality, the Page Rank algorithm and the number of publication.

	Degree Centrality	In-Degree Centrality	Out-Degree Centrality	Betweenness Centrality	Page Rank
1	Graded causation and defaults	Actual causation: a stone soup essay	Necessary and Sufficient Conditions for Actual Root Causes	Graded causation and defaults	Counterfactuals
2	Actual causation: a stone soup essay	Counterfactuals	Explaining actual causation in terms of possible causal processes	Cause without default	Actual causation: a stone soup essay
3	Cause without default	Actual causation and the art of modeling	Causal reasoning in a logic with possible causal process semantics	A modification of the Halpern-Pearl definition of causality	Actual causation and the art of modeling
4	Actual causation and the art of modeling	A Partial Theory of Actual Causation	Causation in Legal and Moral Reasoning	From Programs to Causal Models	A hybrid formal theory of arguments, stories and criminal evidence
5	Actual Causality in a Logical Setting.	Graded causation and defaults	Cause without default	A principled approach to defining actual causation	Representing synonymity in causal logic and in logic programming
6	Counterfactuals	Actual Causality	On Laws and Countefactuals in Causal Reasoning	A general framework for defining and extending actual causation using CP-logic	Embracing events in causal modelling: Interventions and counterfactuals in CP-logic
7	A partial theory of actual causation	A hybrid formal theory of arguments, stories and criminal evidence	Situation Calculus Semantics for Actual Causality	Appropriate Causal Models and the Stability of Causation	Trumping and Contrastive Causation
8	From programs to causal models	Causation: A user's guide	Actual Causality in a Logical Setting.	Actual Causality in a Logical Setting.	Spreading the blame: The allocation of responsibility amongst multiple agents
9	A Modification of the Halpern-Pearl Definition of Causality	Embracing events in causal modelling: Interventions and counterfactuals in CP-logic	Graded causation and defaults	The computational complexity of structure-based causality.	Causal discovery in multiple models from different experiments
10	Explaining actual causation in terms of possible causal processes	A regularity theoretic approach to actual causation	from programs to causal models	Grounding in the image of causation	Discovering cyclic causal models with latent variables: A general SAT-based procedure
11	Necessary and Sufficient Conditions for Actual Root Causes	Actual causation in CP-Logic	Normality and actual causal strength	Normality and actual causal strength	Graded causation and defaults
12	On Laws and Countefactuals in Causal Reasoning	Cause without default	A Modification of the Halpern-Pearl Definition of Causality	Causal analysis for attributing responsibility in legal cases	A partial theory of actual causation
13	Appropriate Causal Models and the Stability of Causation	Interventionist counterfactuals	Arguing about causes in law: a semi-formal framework for causal arguments	Actual causation and the art of modeling	If you'd wiggled A, then B would've changed
14	Situation Calculus Semantics for Actual Causality	If you'd wiggled A, then B would've changed	Probabilistic Reasoning across the Causal Hierarchy	On laws and counterfactuals in causal reasoning	Translating first-order causal theories into answer set programming
15	Causation in legal and moral reasoning	Spreading the blame: The allocation of responsibility amongst multiple agents	Appropriate Causal Models and the Stability of Causation	a A proposed probabilistic extension of the Halpern and Pearl definition of 'actual cause'	A regularity theoretic approach to actual causation

Table 2.5: Top 15 publications according to the Degree Centrality, the In- and Out-Degree Centrality, the Betweenness Centrality and the Page Rank algorithm.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

of Causal Arguments in Law: The Case of Overdetermination", "Explaining Actual Causation via Reasoning about Actions and Change", "Probabilistic Reasoning across the Causal Hierarchy" and "Arguing about Causes in Law: A Semi-formal Framework for Causal Arguments" are added to the set. Finally, after removing books from this set, i.e. removing "Counterfactuals", "Causation: A User's Guide" and "Actual Causality", as well as removing older publications from authors having more than two important publications, i.e. removing older publications from "Denecker", "Halpern", "Hitchcock", "Icard", "Lagnado" and "Vennekens", it contains 36 publications only. [8] Let this set be called $\mathcal{F}$. The graph induced from $\mathcal{F}$ can be observed in Figure 2.14.

To conclude, some literature suggestions are based on the whole graph $\mathcal{G}_f$ irrespective of the relevancy marker. That is, by analysing $\mathcal{G}_f$ it is possible to provide some literature recommendations based on the number of citations a publication has received. Starting with the five articles with the greatest amount of citations, i.e. "Causes and explanations: A structural-model approach. Part I: Causes" [HP05], "Causation" [Lew74], "The intransitivity of causation revealed in equations and graphs" [Hit01], "Structural Equations and Causation" [Hal07] and "Two Concepts of Causation" [Hal04]. Particularly notable is [Lew74], as it is one of the foundational publications responsible for the current surge of interest in the counterfactual approach to causation [BHM09]. Its perceived influence is further supported by the fact that all authors represented in the list above build upon Lewis'es legacy by discussing causation from a counterfactual point of view. However, they differ in their preferred language to represent causal dependencies and in their specific definition of token causality.

Finally, some book recommendations can be given as well. The top five most cited books on the topic of causation are, "Causality: Models, Reasoning and Inference" [Pea09], "Making things happen: A theory of causal explanation" [Woo05], "Causation, prediction, and search" [SGSH00], "Causation, prediction, and search" [SGSH00], "Causation in the Law" [HH59] and "Counterfactuals" [Lew13]. Honourable mentions should be given to the sixth place "Actual Causality" [Hal16a], which provides a great summary of the vast amount of work put forward by Halpern on the topic of causation.

Having defined a set of relevant publications, the next step is to perform a detailed survey of the content put forward in those publications. This can be found in Chapter 3 the objective of which is to use the set of important publications to identify all languages used for encoding causal relationships, all attempts made towards defining token causality and all benchmarks proposed to test the capabilities of said definitions.

[8] [VBD10, BVKPV10, LLLY10, LY10, GDG+10, CH10, GL10, HH11, Shu11, Bri12, Bau13, HHEJ13, HH15, Wes15, CFKL15, BV16, Sch16, Hal16b, BS17, WG17, IKK17, ACHI17, FG17, LG17, Boc18a, II18, BV18, Boc18b, DBV18, BS18, DBV19, LSW19a, LBV19, LSW20, KS20, II20]

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

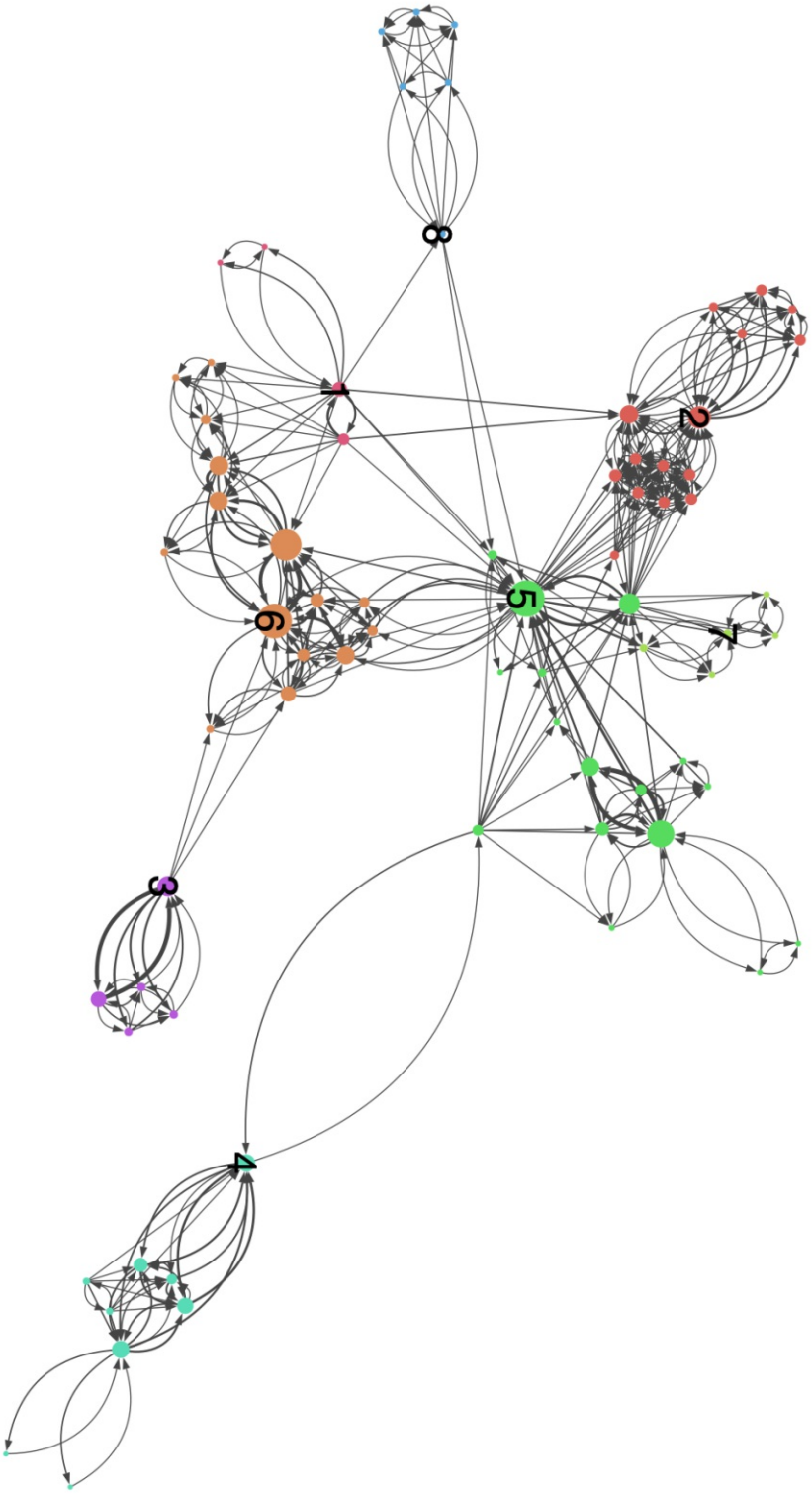


Figure 2.13: A subgraph of $\mathcal{G}_m$, where the colours indicate community affiliation.

TU Bibliothek Your knowledge hub WIEN
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

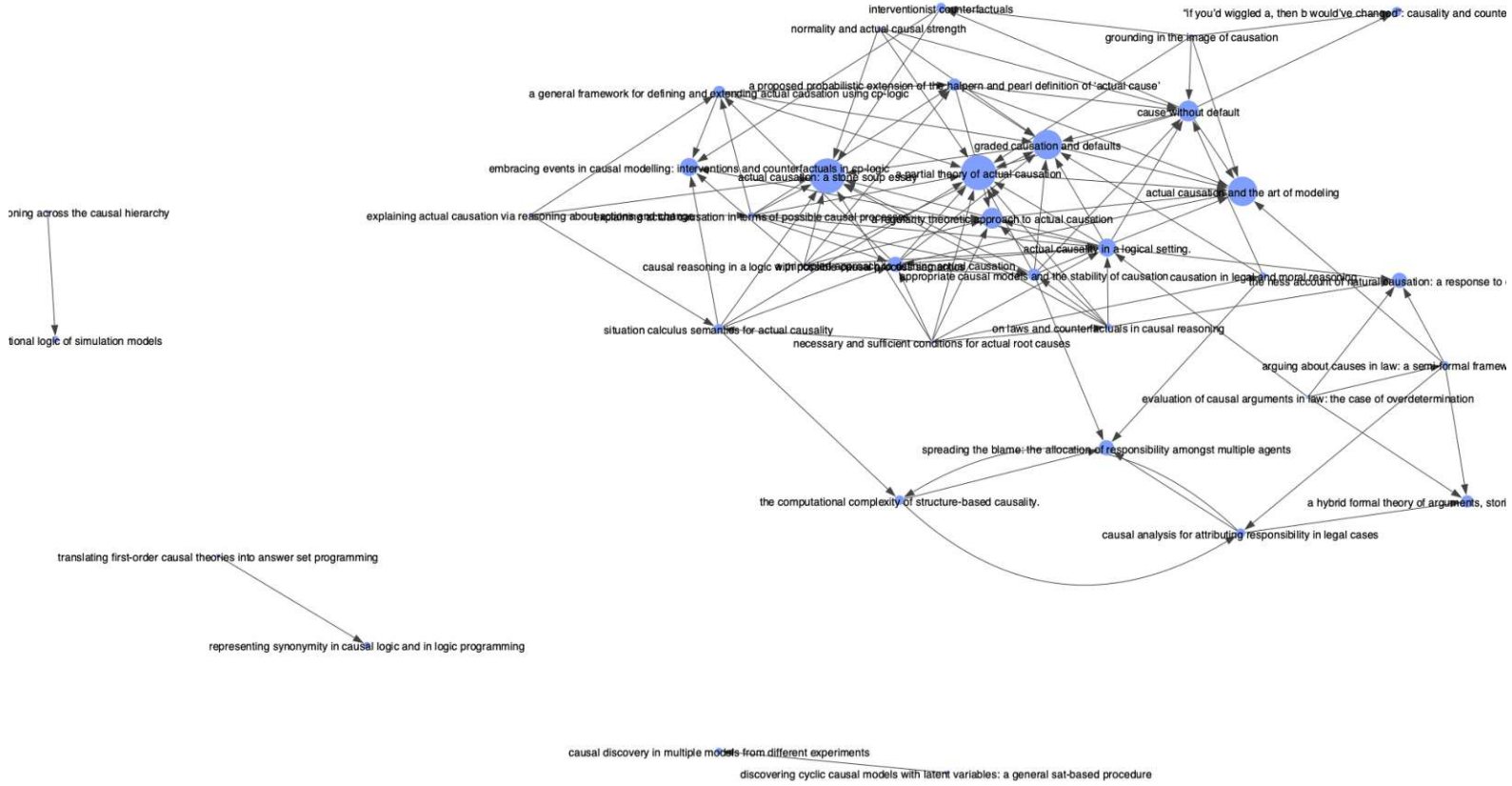


Figure 2.14: The subgraph from $\mathcal{G}_p$ induced by S_A

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

Formalising Causation: A Survey

The quest of formally capturing token causality, brought fourth a plethora of definitions, most of which require some kind of formal representation of causal relationships. May it be equations, rules or mechanisms, the diversity of languages developed in the context of causality is apparent when surveying the publications in $\mathcal{F}$. Although there are some rather popular contenders, the dust has not yet settled on a commonly accepted formal language. Hence, Section 3.1 aims to introduce and roughly classify the diverse range of languages discussed in the publications of $\mathcal{F}$.

The ongoing proposal of new languages, however, is only one side of the coin. That is, the literature contained within $\mathcal{F}$ is littered with definitions that try to capture the elusive concept of token causality from ever so slightly different angles. Being so numerous, Section 3.2 has to be more economical and thus introduces token causal definitions on a rather superficial level, e.g. providing context of its inception, identifying the underlying language and the approach followed in their construction.

It is nice to have languages and definitions, however, their value is drastically reduced, if they fail to satisfy their intended purpose, i.e. capturing token causality. Unfortunately, the whole enterprise of the surveyed literature is to find a formal definition of token causality, thus it is impossible to formally prove whether a particular definition is the "correct one" [Hal]. Hence, authors in the surveyed literature, tend to run their definitions against a battery of benchmarks, to check whether they comply with human intuition. Therefore, to complete the survey contained within this chapter, Section 3.3 will provide an overview of the most frequent examples found in $\mathcal{F}$.

Lastly, it is important to understand that this chapter is intended as an overview, thus it is void of any proper technical definitions and analysis. A more technical inquiry into the subject is postponed to Chapter 4.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.1 Token Causality: Languages

The goal of this section is to convey a rough idea of the available languages, essentially serving as point of reference, guiding the reader to various corners of the literature. This is accomplished by ranking them by popularity and classifying them based on their capabilities, as well as discussing each language family informally.

However, before moving on, some preliminary remarks. This section uses the notion of a language family. This concept intends to express some relation among several languages. Although not rigidly defined family membership is declared either based on the authors own volition, e.g. it is explicitly stated in a publication that a certain language belongs to a particular language family[1], or if the language in question is clearly an extension or a generalisation of another older language.

3.1.1 Overview

When surveying the publications in $\mathcal{F}$ one can detect a total of 9 language families containing roughly 18 individual (semi-)formal languages (i.e. ignoring natural language, abbreviated here with `NL`). However, only two families contain more than one language. In an attempt of gauging the importance of the respective language families, their popularity is measured using the frequency of mentions in the publications from $\mathcal{F}$. Here it is important to point out that by far the most discussed language family is the one building on causal models, with the CP-Logic (Causal and Probabilistic Logic) family and the Non-Monotonic Causal Theory tying for a distant second place and lastly with neuron diagrams taking the third place. To obtain a more fine-grained notion about which publication discusses which language, please consult Table 3.1. Moreover, to obtain a quick overview of the various languages and their properties, please consult Table 3.2.

Starting with the most popular, the causal model family. It contains the greatest amount of languages, making it by far the most developed strain of formalisms. The original, abbreviated here as `CM` and developed in [Pea95], allows for multi-valued variables which are partitioned into exogenous and endogenous variables, a distinction common in the causality literature. Exogenous variables being the variables whose values are determined by factors outside of the model, while the values of endogenous variables are determined by the values of exogenous variables based on the rules relating the variables within the model. The variables are put in relation using a set of structural equations (see Section 3.1.2). While causal dependencies expressed using `CM`, are deterministic and can in theory be cyclic, the literature focuses mostly on acyclic causal dependencies. Furthermore, while being more or less atemporal, there are some implicit temporal aspects emerging from the manner in which causal dependencies are represented.[2] The seminal `CM` was further developed and extended on multiple fronts. For example, the exists `CM+T`, developed in [BV18], which extends `CM` by adding a timing function to make temporal information explicit; `CM+D`, developed in [BS17], which allows one to distinguish between default and

[1] For an example, see [DBV18]

[2] This is particularly relevant for acyclic causal models [BV18]

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

deviant values of variables; CM+N and CM+N2, developed in [Hal08] and [BS17], both of which use a possible world semantics to introduce a notion of normality into causal models. Moreover, there are also attempts to generalise CM to model probabilistic causal dependencies, those are CM+P coined in [FG17] and CM+P2 found in [TK11]. All of the above formalisms retain the distinction between endogenous and exogenous variables.

The second most popular, as well as second most populous family is the CP-Logic family, which contains two independent languages CP and CP2, as well as CP+N which is an extension of CP. The first has its origin in [VDB09], while the second was introduced only relatively recently in [DBV18]. Both formalisms are closely related to logic programming [DBV19]. The third was developed in [BV16] and extends CP by providing the necessary tools for formulating statements using several notions of normality, allowing one to perform normative reasoning within CP+N. To summarise some key properties of the introduced languages. CP is capable of expressing causes on a first-order level, it is designed to model relations between causes and effects in a probabilistic fashion while at the same time allowing them to be cyclic. Moreover, theories within CP are interpreted using a semantic with an explicit temporal dimension. Although CP is already equipped with some sort of default reasoning, CP+N extends CP with additional machinery allowing it to handle normative statements. By contrast, CP2 is solely propositional and atemporal. Moreover, causal dependencies are deterministic as well as acyclic. Yet making a default and deviant distinction, CP2 is capable of default reasoning. Furthermore, all formalisms in the CP-Logic family require one to specify endogenous and exogenous variables.

The publications in $\mathcal{F}$ discuss two slightly different variants of non-monotonic causal theories. However, according to [Boc18a] both versions are equivalent. Hence, they are counted as the same language in this survey. To be precise, CT will henceforth reference the version developed in [Boc03] as it has a higher relevance in the context of token causality (see [Boc18a, Boc18b]). By contrast, within the context of $\mathcal{F}$ the original version, introduced in [MT$^+$97], was never used to develop a definition for token causes. CT is essentially a binary propositional language, extended by an atemporal causal inference relation wrapped in a non-monotonic fixed point semantic. Moreover, while not explicitly discussed in the context of token causality, CT has the machinery necessary for default reasoning. Furthermore, diverging from the previous formalisms no distinction between endogenous and exogenous variables is made.

The original form of neuron diagrams was, according to [Hit07b], first introduced in [Lew86]. The size of the language family surrounding neuron diagrams can unfortunately not be properly gauged. This is because in most cases neuron diagrams are only introduced and used in an informal manner [EW10]. Hence, this thesis uses ND to reference the general family of neuron diagrams, rather than to a specific language. ND differs drastically from other formalisms discussed in this survey. The main reason supporting this judgement is that ND is (in practice)[3] a purely graphical formalism. Unfortunately, it is quite difficult to assess the capabilities of ND. However, they do not allow for cyclic dependencies

[3] [EW10] provided a definition of neuron diagrams that does not rely on a graphical representation.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

among causes and their effects and they usually consider those relations as deterministic. Additionally, ND distinguishes commonly between endogenous and exogenous variables. Similar to CM, ND is not a formalism that explicitly deals with time. However, ND diverges from CM, by distinguishing between default and deviant values, i.e. ND provides some form of default reasoning.

Apart from the top most commonly discussed languages, $\mathcal{F}$ contains a plethora of other formalisms. Firstly, the language called situation calculus. It has its origin in [MH69]. However, it seems that there are several variants of situation calculus. Nevertheless, within the publication in $\mathcal{F}$, only [BS18] and [KS20] are using situation calculus to model causal dependencies. As the variants used in both are identical, it is the only one discussed and thus is henceforth referenced as SC. Secondly, there exists rather new and unique formalism based on so-called simulation models while relying heavily on conditional logic and Turing machines. This formalism is introduced in [II18] and referenced here with SM. Thirdly, there is the action language AL, introduced in [BG00] and re-purposed for the application in the realm of causality by [LBV19]. Fourthly, [LSW20] introduced a domain specific causal language designed for modelling arguments in legal argumentation and liability disputes. They called it "(semi-)formal framework for causal argumentation", abbreviated here with FCA. Fifthly, [Bau13] extended classical first-order logic allowing him to express type causal relationships, his approach is abbreviated with FOL+. Lastly, there is causal abductive reasoning, referenced here using CAR. [BVKPV10] used CAR in conjunction with another formalism concerned with modelling argumentation to build an extensive framework for establishing facts in legal cases.

3.1.2 Causal Models

The original causal model formalism CM was spearheaded by Judea Pearl. Being the most popular formalism, it is discussed by a plethora of different authors making it fairly unique among the formalisms captured in this survey. The assumption present in most members of this family is that the causal mechanisms governing the world can be described by a set of (random) variables and a set of deterministic structural equations[4] [Hal15b].

A structural equation allows one to condense all type-causal relations that may influence a variable into a single equation. Those equations are not algebraic and are best understood as assignments, i.e. they fail to be symmetric. A rather sensible choice, as a cause influences its effect, while an effect does not necessarily impact its cause. For example, a volcanic eruption may cause one to reconsider their plans of going on vacation in Pompeii, yet not taking a vacation in Pompeii is (most likely) not a cause of said volcanic eruption. Although causal models can have cyclic relationships among their variables, so called acyclic causal models tend to be the primary subject of investigation. Intuitively, a causal model is considered acyclic, if one can order the endogenous variables, such that the variables lower in this order are independent of the above them. Moreover,

[4] The concept of structural equation emerged according to [BS18] in [Sim55]

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Articles	CM +*	CP/CP2	SC	CT	FOL+	ND	SM	AL	FCA	CAR
[VBD10]		✓								
[BVKPV10]										✓
[LLLY10]				✓						
[LY10]				✓						
[GDG+10]	✓					✓				
[CH10]	✓									
[GL10]	✓									
[HH11]	✓									
[Shu11]	✓									
[Bri12]	✓									
[Bau13]					✓	✓				
[HHEJ13]	✓									
[HH15]	✓									
[Wes15]	✓									
[CFKL15]	✓									
[BV16]		✓				✓				
[Sch16]	✓									
[Hal16b]	✓									
[BS17]	✓									
[WG17]										
[IKK17]										
[ACHI17]	✓									
[FG17]										
[LG17]	✓									
[Boc18a]	✓			✓						
[II18]							✓			
[BV18]	✓									
[Boc18b]	✓			✓						
[DBV18]		✓								
[BS18]	✓		✓							
[DBV19]	✓	✓								
[LSW19a]									✓	
[LBV19]								✓		
[LSW20]	✓								✓	
[KS20]	✓		✓							
[II20]	✓						✓			

Table 3.1: Depicts which publication discuss which languages families

Formalism	Year	Quantification	Multi-valued Variables	Default Reasoning	Temporal	Probabilistic	Origin
SC	1969	✓			~		[MH69]
ND	1986			~			[Lew86]
CAR	1993			✓			?
CM	1995		✓			~	[Pea95]
CT	1997			~			[MT+97],[Boc03]
AL	2000				~		[BG00]
CM+N	2008		✓	✓			[Hal08]
CP	2009	✓			✓	✓	[VDB09]
CM+P2	2011	?	?	?	?	✓	[TK11]
FOL+	2013	✓		✓			[Bau13]
CP+N	2016	✓		✓	✓	✓	[BV16]
CM+D	2017		✓	~			[BS17]
CM+N2	2017		✓	✓			[BS17]
CM+P	2017		✓			✓	[FG17]
CM+T	2018			~	✓		[BV18]
CP2	2018			~			[DBV18]
SM	2018						[II18]
FCA	2020			✓			[LSW20]

Table 3.2: Summary of languages found in $\mathcal{F}$. This table categorises the languages based on whether they allow for quantification over variables; whether the variables used are multi-valued; whether the language is equipped with some form of default reasoning or is capable of encoding statements about normality; whether the language makes time explicit or whether it particularly emphasises sequences of events; whether the causal relations (or variables) can be probabilistic. (A check mark means that it fully satisfies the property; A tilde means that it partially satisfies the property; A question mark means that it is unknown if property is satisfied)

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

acyclic models allow for simpler reasoning, this is because in such models the value of all endogenous variables can be uniquely determined given any context, i.e. a value assignment of the exogenous variables. By contrast, cyclic models could have several solutions to the set of structural equations, i.e. there exists more than one fixed point in the computation [Hal15b].

Since structural equations encode all possible causal relations between variables it is possible to perform interventions on them. Intuitively, performing such an intervention is akin to asking the question "What would happen, if I change the value of variable X from x to value x'?". Technically, one intervenes on a causal model by replacing existing structural equations with fixed values or different structural equations and by recomputing the solutions for the set of structural equations. Another peculiarity of structural equations is that they are deterministic. Initially, this may strike as a rather strange decision, as it is tempting to declare the relationship between cause and effect to be probabilistic, e.g. a lightning strike has a 60% chance of causing a forest fire. However, a deliberate choice was made to diverge from inherently probabilistic approaches such as Causal Bayesian Networks, by keeping structural equations deterministic. [Pea09] justifies this by drawing an analogy to the Laplacian and the quantum mechanical conception of physics. That is, the former considers natures laws as deterministic with uncertainty only emerging due to ignorance, while the latter understands determinism as a mere approximation of inherently probabilistic laws. However, as the goal of this endeavour is to capture and model the intuitive human-level understanding of causality, the focus on the former is apt [Pea09]. Similarly to Pearl, Halpern advises against probabilistic causal dependencies, suggesting instead the expansion of the model such that this uncertainty can be described, e.g. adding variables such as dryness or altitude. However, as this is not always possible, one can easily push probability out of the equations by putting a probability distribution over the exogenous variable in a causal model [Hal15b, Hal16a, p. 13].

Over the years, causal models have been criticised by many. Neglecting criticism about the lack of generalisation to a predicate level raised in [BS18], most of its critiques have lead to the creation of independent extensions.

Firstly, to overcome the confinement to deterministic structural equations, [FG17] proposed `CM+P` as a probabilistic extension of causal models. However, this is only the most recent of such attempts, with at least another one, i.e. `CM+P2`, detected in the literature. This indicates that irrespective of Pearl's arguments, seem to be some potential applications and thus some desire for a purely probabilistic causal framework. One particular advantage of `CM+P` and others of its kind would be a better resonance with common scientific theories, most of which are probabilistic in nature [FG17].

Secondly, [BV18] advocate for extending causal models by a timing function, which essentially allows one to incorporate temporal information directly into the causal model. Hence, they create `CM+T`, which is better suited to deal with those problematic examples in the literature that derive their difficulty from requiring temporal precision. By contrast, if one wants to model time within `CM`, one is required to simulate it by adding timestamps

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

to the variables in a causal model. While cumbersome, [II18] argues that requiring explicit temporal information during the modelling process, is too stringent especially since there are cases where causal relationships do not require such an information.[5]

Finally, a significant thrust behind the desire to improve upon CM, is a group of examples, called non-structural counterexamples, that suggest that isomorphic causal models can have differing causal intuitions. Although contested, see [BS17], it is the perceived failure of causal models on that front, that motivated the extension of causal models with some theory of normality, e.g. CM+D and CM+N. CM+N requires the modeller to rank the various contexts based on their perceived normality, i.e. CM+N operates on a possible world semantic. This extension equips the modeller with a high degree of flexibility when constructing a causal model. On the one hand, this provides certain advantages, such as allowing one to capture the distinction between conditions and causes present in the legal tradition, with conditions being values of variables with a higher degree of normality. While on the other hand, it induces a host of other problems, e.g. it provides the modeller with the ability to hard-code their desired token causes into the model. This further exacerbates Hall's complaints about structural models. That is, he argues that the structural equations approach places a much greater emphasis on problem modelling, amounting to little more than building the solution into the model [BS17, HH15, Wes15, EW10].

3.1.3 CP-Logic

The language CP, seems to be the first member of the CP-Logic family. It was developed in [VDB09] as a probabilistic logic programming language capable of expressing probabilistic causal laws with an informal semantics independent from the epistemic agent-based semantics of deterministic logic programs and the frequentist interpretation present in the probability calculus. They view probabilistic causal laws as follows. Each law connects a cause with its possible effects. That is, an event[6] has multiple possible effects. If such an event occurs, based on the probability indicated by the rule, only one of the possible effects will be realised. On the semantic side, they took inspiration from action languages and used an approach championed by Shafer to create an appropriate semantic for this language. In particular, the underlying idea is that causal and probabilistic concepts should be evaluated dynamically, i.e. they should be understood as a story explaining how the domain evolves. Shafer formalises this intuition by relying on probability trees. In such a tree, vertices represent states and edges represent events that induce a state transition and are labelled with probabilities.

[5] Although in this particular case CM+T allows one specify the timing only partially, which somewhat weakens this argument.

[6] Due to the fact that this language operates in the intersection of both logic programs and probability theory, they have to distinguish, between events that cause transitions between states and events that are set a collection of possible outcomes. Hence, they follow Shafer and call the former Humean events and the latter Demoivrian events

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

In [VDB09] they contrast their language against several other languages. Most notably, CM and CT. Firstly, CP is more expressive, as well as seemingly better suited for modelling cyclic causal relations than CM. Hence, rather than focusing solely on acyclic causal relations, most of the discussion surrounding CP omits this restriction. Another supposed benefit of CP is that it alleviates one from the requirement of compressing all causal influences on a variable into a single structural equation. Secondly, CP differentiates itself from CT by having an initial state, at which all variables are in their default state, which changes during the evaluation process. Hence, CP is capable of distinguishing between default and deviant values, requiring explanation only for the latter. By contrast, the original semantics of CT is a fixed point semantics that requires an explanation for any variable value. Put differently, CT treats truth and falsity symmetrically, by requiring a causal explanation for both positive and negative occurrences of propositions, while in CP truth and falsity are treated asymmetrically due to the fact that only the deviation from the natural state of a variable must be causally explained.

Moreover, due to its constructive nature CP rules out any unfounded causes, i.e. causes that can cause themselves, a property not satisfied by CT. One repercussion of this is the inability of CT to express cyclic causes. Unfortunately, this cannot be avoided, as the rule structure responsible for unfounded causes is required to introduce exogenous variables into a theory. A feature not required in CP as this distinction is already made explicit in this language. To summarise the advantages of CP are the probabilistic component in the endogenous part of the model, the ability to encode default and deviant variables and its temporal semantic [BV16, VDB09].

However, CP is not without criticism. Firstly, since a theory in CP is defined as a finite set of laws, the language is naturally restricted in its expressibility, e.g. one would be unable to express a scenario where a die is rolled as long as it takes to obtain a six. Secondly, all outcomes of an event must be known during the modelling process. Thirdly, it is impossible to model events that cancel out or reinforce each other's effects, a defect somewhat remedied in CP2. Fourthly, the language is ill-equipped to speak of contributing causes, e.g. turning on a tap would not instantaneously cause a basin to be full, but only contributes a certain amount per time unit [VDB09]. Fifthly, [BS18] criticises the expressivity of CP on the grounds that it is unable to distinguish between properties and actions, as well as the non-existence of quantified effects. Lastly, similar to the extension of CM, CP was extended to CP+N in [BV15], motivated by the desire to produce normative statements within the CP framework. Without going into details, they leverage the probabilistic nature of CP to distinguish between statistical and normative normality. Both of which are implemented by operations on the possible set of effects triggered by an event. This equates to removing certain effects from CP-rules depending on the type of normality and can be envisioned as refining the probability tree based on how normal each branch is.

The language CP2, seems to be an entirely separate language, declared to be within the same family as CP. CP2 is a language that allows one to model causal processes by representing the underlying causal mechanisms of a situation. Secondly, taking

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

inspiration from neuron diagrams, the authors argue for the necessity of distinguishing between conditions that trigger other causal mechanisms and conditions that allow for the preemption of such mechanisms. This distinction relates CP2 to non-monotonic formalisms, such as Default Logic [Rei80] or Answer Set Programming (see [BET11]).

On a syntactic level, a causal theory in CP2 is a set of causal mechanisms, with each causal mechanism being described by a set of triggering conditions, a set of enabling conditions and an effect. The triggering conditions set the mechanism in motion that produces the effect, if all enabling conditions are met. In stark contrast with CP, theories in CP2 are a non-empty sets of deterministic causal mechanisms without any cyclic causal dependencies.[7] Although, according to [DBV19] the language can easily be extended to allow for cyclic causal relationships. An additional oddity of such theories is that they cannot contain contradictory mechanisms, i.e. one mechanism producing an effect and another producing the negation of this effect [DBV19].

From a semantic perspective, CP2 not only implements the law on inertia, i.e. a change in state requires an external force, but also adheres to Leibniz's principle of sufficient reason, i.e. every true fact has a reason. This is accomplished by relying heavily on the default-deviant distinction. To be more precise, for every endogenous proposition it is assumed that the causal theory contains all causal mechanisms affecting it. Furthermore, each endogenous proposition has either a default or deviant state. Since a causal theory can only contain non-contradictory mechanisms, mechanisms can only have deviant effects. Now, with every proposition being in its default state, a causal process starts by firing applicable but unsatisfied causal mechanisms, i.e. mechanisms with satisfied conditions but unsatisfied effect. The firing of this mechanism forces the proposition to take on its deviant state. The process continues until all causal mechanisms are satisfied. Notice that due to the structure of a causal theory, a proposition can change its value at most once. Hence, any proposition remaining in its default state is true by inertia. Moreover, while deviant states are justified by the rule causing the state switch, default states are justified by the fact that no rules with the appropriate effect are satisfied [DBV19].

Due to the theories in CP2 being non-contradictory sets of causal mechanisms, the language can not handle scenarios with alternating variables, e.g. it is unable to model a simple light switch. In [DBV19] additional work and open questions are surrounding this language are mentioned. That is, they express the need to develop extensions that allow for the modelling of probabilistic and cyclic causation. Furthermore, to capture a greater array of real-world problems elevating this language to the first-order level is required. Moreover, the language remains unexplored from both the computational complexity, as well as the proof theoretical perspective.

[7] Unfortunately, it is not specified, whether such theories must be finite or can be infinite in size.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.1.4 Non-Monotonic Causal Theory

CT was originally conceived as a non-monotonic formalism for reasoning about action and change in AI in an attempt to deal with the frame problem, see [GLL+04]. At its core, it tries to adhere to Leibniz's principle of sufficient reason, as well Pearl's claim that relevant situations are determined not only by the rules that belong to the causal theory, but also by what does not belong to it. The latter is accomplished by using the non-monotonicity of CT.

Syntactically CT is fairly simple. That is, it extends an ordinary propositional language with a causal relation, which expresses that a proposition causes another proposition. This causal relation can be defined in several ways. However, in [Boc18a] a modified inference relation taken from the input-output logic described in [MVDT00a] was chosen. This inference relation is a production inference relation, a defining feature of which is its failure to satisfy the reflexivity postulate, i.e. a proposition cannot cause itself. However, in order to arrive at a causal inference relation, [Boc18a] strengthens the relation further, such that it is fairly similar to classical entailment, but for the fact that both the reflexivity and contraposition remain unsatisfied. On top of the logic obtained by extending propositional logic with the described relation, resides a non-monotonic semantics. This semantic ensures that only those models that are closed w.r.t. the causal inference relation are accepted. Resulting in every proposition being causally explained.

As noted in [Boc18a] CT has some glaring representational deficits. The first is that CT uses only binary variables and is therefore unable to encode situations that are easily modelled by CM. The second is its inability of expressing notions of normality or perform any other form of default reasoning. However, [Boc18a] claims that the formalism originally developed in [MT+97], is well equipped to remedy those deficits. Moreover, the causal relation introduced is atemporal, thus similar to causal models, if one deals with time-critical scenarios it is necessary to marking propositions with time stamps. Lastly, it is one of the few languages within causality literature that does not explicitly distinguish between endogenous and exogenous variables [Boc18a].

3.1.5 Neuron Diagrams

In their simplest form, a neuron diagram can be understood as a directed acyclic graph, where each vertex, called neuron, can either fire or not, often indicated by its colour. A neuron can be either exogenous, i.e. it has no incoming edges, or endogenous, i.e. it has at least one incoming edge. Moreover, edges between neurons can also be separated into two categories, i.e. stimulating edges and inhibiting edges, often distinguished through having a triangle and respectively a circle as arrowhead. In its simplest form, such neuron diagrams follow a fairly straightforward semantic. That is, while it is externally specified whether an exogenous neuron fires or not, an endogenous neuron fires if and only if it is stimulated by at least one firing neuron, and inhibited by zero firing neurons [Hit09, EW10, Bau13].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Unfortunately using only one kind of neuron is insufficient to capture many examples, e.g. encoding a conjunction is already difficult. Hence, both [Hit09] and [Bau13] define and utilise alternative, more complicated neurons. For example, one could consider a stubborn neuron that only fires, if all or some of its predecessors fire as well, or a neuron that only fires if the number of stimulating inputs is greater than the number of inhibiting inputs.

Using neuron diagrams as a formalism to encode causal structures is particularly ubiquitous in the philosophical literature. Due to their graphical nature, they provide a rather intuitive method of representation of causal dependencies for the small scale examples common in the literature. This simplicity naturally restricts this language in its expressivity, e.g. in their common form they cannot encode causation by omission. Their use was criticised in [Hit09] on similar grounds. That is, it is their failure to encode complex relationships between variables, that makes him an advocate for the use of structural equations, as for example used in causal models. Although acknowledged in [EW10] they justify the use of neuron diagrams by citing their simplicity. In opposition to Hitchcock, Hall criticises the structural equation approach in [Hal07]. That is, while acknowledging their value, he perceives their status as inflated, favouring neuron diagrams instead. In particular, he endorses them not only due to their simplicity, but also due to their ability to encode a default/deviant distinction. [Bau13, EW10, BV16].

3.1.6 Situation Calculus

Situation Calculus has a long history in the causation literature. However, the primary purpose of situation calculus is the modelling and reasoning about dynamic systems, whereas formalising causality was only a secondary objective. During the 2000s the causality literature experienced a surge of interest in the situation calculus. At that time Judea Pearl tried to use situation calculus to remedy some deficits with the causal model approach. In particular they wanted to rectify the failure of CM to distinguish between transitional and enduring conditions [VBD10, KS20, BS18].

Being a rather involved formalism, describing it in an informal manner is bound to produce a rather muddied picture. Hence, the subsequent paragraph tries to convey a rough intuition only. SC is a many-sorted situation calculus variant that is used to define a basic action theory (BAT). According to [KS20] and [BS18] the basic action theory approach was developed by [Rei01]. A BAT has several components.

Firstly, action terms. Such terms represent actions that can be performed in the modelled system, e.g. a car taking a left turn. Moreover, in a BAT there must be a set of action precondition axioms that specify the preconditions of a situation that are required for the execution of a certain action. Secondly, situation terms. A BAT requires that the initial situation of a system must be specified by using a set of initial state axioms, e.g. at which intersection a car may be. From there, any other (complex) situation term consists

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

of a sequence of action terms[8] and the initial situation, e.g. at which intersection a car may be after taking a left and a right turn. The language is equipped with a predicate that encodes the partial ordering of those situation terms, thus allowing one to checks whether one situation can be reached from another by performing a sequence of actions. Thirdly, fluents. In a BAT fluents are situation dependent relations, e.g. a relation that check whether a specific car is at a certain intersection in a particular situation. Such fluents are subjected to a set of successor state axioms, those specify after what actions and in which situations a fluent can change its value. It must be noted that in a BAT it is permitted to use non-fluent relations that are situation independent relations, e.g. to encode the layout of a street network. Lastly, a BAT must also contain additional axioms, e.g. unique name axioms.

Of the discussed languages `SC` is deemed to be on the more expressive side, so much so, that it is possible to produce a token causal definition that can identify causes of conditions expressed in first-order logic. Additionally, `SC` is equipped with a plethora of features, that allow for the modelling of actions. However, according to [BS18] `SC` still lacks in expressiveness, identifying the need for language that can model time explicitly and that allow for concurrent actions, i.e. a situation calculus variant where actions are only partially and not totally ordered. Although contested in [BS18], relying on some form of situation calculus to model causality has been criticised in [VBD10]. They argue that its heavy machinery is not entirely necessary in the context of causality, i.e. they speak of it being an "overkill" [BS18, KS20].

3.1.7 Simulation Models

`SM` take a rather unorthodox view on modelling causality. At its core, this language is founded on the belief that causal (and conditional) reasoning is closely related to simulations. A belief that is supported by some empirical evidence, see [Jel07]. Expressed differently, in order to answer a question such as "what would have been if?", humans simulate the hypothetical scenario in their head. Inspired by that this language provides a flexible framework for evaluating the truth values of propositions, by simulating conditionals using simulation programs, which opens up the possibility of incorporating generative models developed using deep neural networks into the reasoning process.

Syntactically the major feature of the conditional logic underlying simulation models is the ability to express interventions, i.e. one can construct sentences such as "If A were true, then B would be true". In [II18], interventions can only be expressed as conjuncts of propositions, while all remaining formulas have the full toolset of propositional logic available. Sentences of this language can be evaluated using causal simulation models, they consist of a Turing machine and a start tape, encoding the truth values of each

[8] If the sequence of action terms is variable free, such a sequence is called a narrative. Since the literature in the intersection between law and causality heavily emphasises the importance of stories, there may be the potential for applying the concepts discussed in [BS18] and [KS20] within the field of legal reasoning.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

propositional variable. A statement containing an intervention evaluates to true if the proposition that should hold after the intervention is true on all halting executions of the Turing machine where the values of variables that were specified in the intervention remained the same value as specified in the intervention.

Some relevant properties of this approach are that it does not require temporal information when modelling causal scenarios, as according to [II18] requiring temporal information always be made explicit is too stringent. A similar view is also present in the design of the classical causal models. While being similar in one regard, according to [II18] those two approaches deviate drastically on one of the fundamental principles of conditional reasoning, namely cautious monotonicity.

There are several possible pathways for improving this language. The first is to elevate SM to the first-order level, allowing one to perform interventions on a first-order basis. The second is to introduce probability into the system, and the last would be to impose orderings on the variables, e.g. through timestamps, in order to provide a language akin to acyclic causal models, i.e. causal models with no cyclic causal dependencies.

3.1.8 Other

Here the remaining less prominent, as well as less formal languages, i.e. FCA, AL, FOL+ and CAR, are discussed. FCA is the only domain-specific causal language discussed in this thesis. Since it was designed for the analysis of causal arguments in legal argumentation and liability disputes, it is deliberatively designed to be semi-formal. That is, while [LSW20] acknowledges the benefits of fully formalised languages such as CM, they express the concern that the technical details present in such fully formalised languages discourage their application in law. FCA contains three basic structures. That is, factual propositions, predicates for similarity, evidentially and causal links, as well as a set of inference rules. The similarity and evidentially predicates, allow one to express the "similarity" between propositions, as well as the existence of evidence for a certain proposition. The causal link predicate is constructed as a ternary relation, two positions are used to link cause and effect, while the last position indicates the certainty of the causal link, i.e. it determines whether the specified cause normally or always produces the effect. This is precisely the reason why, FCA has two inference rules, one being classical and the other one being defensible. The meaning of those rules is conveyed using a set of rule schemata, expressing properties such as "If there is evidence for propositions, those propositions hold" or "Similar propositions cause the same effect".

AL by [BG00] is part of a class of action languages, developed for reasoning about actions and their effects and has close ties to non-monotonic formalisms such as Answer Set Programming. In particular, this language has the ability to represent both direct and indirect effects of actions. According to [LBV19] languages such as AL are required for understanding token causality, as focusing on actions and their direct (as well as indirect) effects allow for a deeper understanding of causal mechanisms [LBV19]. Semantically

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

AL relies on transition diagrams, consisting of a collection of states and a set of triples representing state transition induced by events. This language distinguishes between events and fluents, which are propositions that can change in value over time. There are three kinds of statements in AL. The first are called dynamic causal laws and express that in the case of an event, given that all conditions (expressed as literals) hold, the consequence of this law holds in the next state. The second are called state constraints, one can use those to express that any state that satisfies the conditions of this rule must also satisfy its consequent. The last, being executability conditions prevent events from occurring, in case that the specified conditions hold. Similar to the SC, AL is also a language designed for the modelling of action and change. However, according to [LBV19], AL is better suited for the representation of indirect effects of actions, a property that seems to be rather desirable [LBV19, Boc18a][9]. In another comparison with SC, [LBV19] highlights the simpler semantic of AL over SC, as AL does not rely on first-order logic.

FOL+ relies on a light analytical toolbox, only employing material conditionals, standard Boolean minimization procedures, and an additional stability condition. At its core, FOL+ uses first-order logic to express type causal relations. In particular, it uses predicates to identify objects in the domain as events of a particular kind, allowing one to formulate statements such as "an event of type A, causes an event of type E", with the causal relation being merely a shorthand for a set of first-order sentences with equality. However, in order to evaluate such sentences appropriately, [Bau13] develops the notion of a minimal theory, specifying what are minimally sufficient and necessary conditions in order to interpret a material regularity (expressed using material implication) as causal. Moreover, akin to many other approaches, [Bau13] proposes a possible extension of FOL+, to account for notions of typicality and normality, i.e. similar to CM+N, this is accomplished by ranking the variable assignments.

CAR is, as indicated in its name, related to other abductive model-based reasoning approaches. In general, those approaches are a form of non-monotonic reasoning, that try to find plausible explanations for the state of the world, with respect to an inference relation, e.g. material implication [Pau93]. In particular, CAR is a formalism that operates on the propositional level and tries to establish causes from hypothesis and causal rules. An inference in CAR is called a story, those are defined similar to derivation in classical logic, i.e. a story is a sequence of propositions where each proposition is either a hypothesis or derivable from earlier propositions in the sequence using some causal rule. The emphasis on stories in the introduction of the semantics of CAR reflects the story-bases approaches present in the legal tradition. Moreover, the formal notion of a story, also adheres to some of the properties desirable in the informal story-based approaches. In particular, the requirement of a story to be internally consistent and the need to be plausible, i.e.

[9]Moreover, in [Boc18a] the following is stated: "According to Pearl, causal assumptions are encoded in the missing links (that sanction, e.g., claims of zero covariance)."

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the story must conform with the knowledge about the world. In [BVKPV10] CAR was used in conjunction with another formalism designed for modelling argumentation, to build a more comprehensive formalism for studying establish facts in legal cases, e.g. criminal cases [BVKPV10].

3.2 Token Causality: Definitions

The languages introduced may enable the modelling of causal dependencies. However, this alone is not sufficient for identifying token causes. For example, the structural equations in an acyclic causal model allow one to encode causal relationships on the type level. Using those equations and a description of the world it is possible to determine the precise value of each variable in the model [Hal15b]. Clearly, this mode of inference is forward-looking, and is thus in stark contrast with the reasoning employed in the context of token causality, which is characterised by a backwards looking mode of inference. That is, the main objective is to identify a suitable set of variables that explain why a particular variable has a specific value. Finding an appropriate definition for this kind of causality is quite the undertaking, as indicated by the lively debate surrounding this subject, i.e. given 9 language families containing 18 formal languages, a total of 32 definition are discussed in the articles from $\mathcal{F}$.[10] Unfortunately, not a single one, is sufficiently precise to satisfy all the toy examples found in the literature without contention. However, the linage of token causal definitions developed by Halpern is often used to benchmark new definitions, making it the closest the literature has to offer to an accepted standard.

The section unfolds as follows. Firstly, the definitions build on CM will be discussed. Secondly, all definitions relying on formalisms other than CM are presented. Thirdly, the definitions found using semi-formal or informal languages are highlighted. Lastly, the section concludes with the categorisation of the definitions introduced above. To reiterate, all of the above introductions, refrain from engaging with technicalities. Such content can be found in Chapter 4 for a small selection of the definitions discussed here.

3.2.1 Definitions based on Causal Models

Starting with the definition formulated in the most popular formalism CM. The chronologically first definition, referenced as HP-01, is due to Halpern and Pearl (HP). It was originally formulated in [HP01], inspired by Pearl's notion of causal beam (see [Pea98]) and uses counterfactuals to identify token causes. Shortly after [HP03] proposed an example, that seemingly demonstrated that HP-01 is insufficient. Leading to the creation of the updated HP-definition, abbreviated as HP-05. It originated in [HP05] and is by far the most popular, i.e. the most widely used and discussed, definition yet. Being merely an update of the original, HP-05 remains firmly rooted in the counterfactual tradition. Unfortunately, it was demonstrated in [ACHI17] that the computational complexity of

[10] Four of the 32 definitions, i.e. But-For, INUS, NESS and "causally relevant factor", tend to be defined using natural language.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

finding causes with HP−05 is D_2^P-complete[11] for both binary and general causal models. By contrast, [EL02] demonstrated that HP−01 is merely NP-complete in the binary and Σ_2^P-complete in the general case. Fortunately, the necessity of HP−05 was challenged in [Hal16b], where it is argued that the model used to discredit HP−01 neglected to properly formalise the provided scenario. This was followed up by demonstrating that with a small, but unfortunately not always natural, expansion of said model HP−01 will produce judgements similar to HP−05. In [Hal15a] an new variant of this family was formulated. This definition is referred to as the modified HP definition and is abbreviated here with HP−15. According to Halpern this definition is not only conceptually and computationally simpler, but also provides the more preferable answers. That is, it deals with the critique raised in [HP03] and it handles various examples better than HP−05, e.g. Hall's non-existent threat example [Hal15a, Hal16a, Hal07, p. 27]. With respect to computational complexity, HP−15 is NP-complete in the binary and D_1^P-complete in the general case, making it the most efficient HP-definition yet [Hal16a, p. 153-154]. In [FG17] a variation of HP−05, namely HP−05c, was presented. The purpose of which was to adjust HP−05 to make contrastive causal judgements. This extension is based on the view that causation is contrastive in nature, thus it is not a binary, but a tertiary relation. For example, administering one dose of medicine saves the patients life and administering a second dose is absolutely redundant, i.e. same outcome as giving only a single dose. Hence, giving two doses instead of zero caused the patient to survive, while giving two doses instead of one is immaterial to the person's survival.

There are also independent definitions that use only the basic causal model variant CM. Firstly, there is Hitch−01 which was provided in [Hit01]. Secondly, there is Wood−03 formulated in [Woo05]. Thirdly, in [GDG+10] two simplified versions of HP−05 are proposed, abbreviated here with Simple and SimpleJ. However, all of the above disagree with HP−05 on some of the traditional examples found in the literature. Lastly and particularly of note is the "Partial Theory of Actual Causation", or abbreviated PTC, put forward in [Wes15]. He claims that his version improves upon the HP−05 definition. This is partially accomplished by incorporating some tools from the regularity theoretic tradition without extending the causal model by some additional structure.

Although subject of contention, e.g. see [BS17], over the years HP−05 was extended by some form of default reasoning, i.e. they incorporated a normality ordering over various contexts. This extension was motivated by the discovery of several so-called non-structural examples, i.e. examples that have the same causal model and yet exhibit different intuitive answers. Hence, it using such extended causal models provides the necessary flexibility, to resolve the issues put forward by those examples. Additionally, this notion of normality brings forth the possibility of introducing normative reasoning into causal judgements. One such definition found in the selected literature is HP−05d, which is a definition that extends HP−05 with default reasoning. However, as mentioned

[11] D_k^P is defined in [ACHI17] as the set of all languages L_3 such that there exists a language $L_1 \in \Sigma_k^P$ and a language $L_2 \in \Pi_k^P$ such that $L_3 = L_1 \cap L_2$. Σ_k^P and Π_k^P are simply levels on the polynomial hierarchy (see [AB09, p. 97-99]). This complexity class is a generalised version of the $k = 1$ case coined in [PY82]

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

in [Hal16a, p. 97-103], it is possible to extend HP−15 in a similar fashion. Apart from the HP-definition, there are other definitions that use causal models and incorporate some notion of normality. In [BS17] the definitions HmM and MbM can be found. The first one is called *Hitchcock-meets-Menzies* and is a modified version of a definition found in [Hit07a]. It incorporates normality assumptions by simply partitioning the range of each variable in a causal model into default and deviant values. The second one is called *Menzies-by-Menzies*, it builds on the ideas proposed in [M$^+$07] and requires one to rank the possible states of the world based on their normality. Hence, this approach can be seen as being similar to HP−05d.

Another extension of HP−05 can be found in [FG17], where they generalise HP−05, which is solely concerned with deterministic cases, to the probabilistic case. While not demonstrated in the article in question, they conjecture that their natural probabilistic extension is set up to deal with a wide range of examples circulating in the literature. Furthermore, they claim that they improved upon a previous attempt by [TK11] where they tried to generalise causal models to the probabilistic level using "Probabilistic Active Paths", thus their approach is abbreviated here with PAP.

Lastly, there exists a definition that, while relying on some CM-variant diverges from the HP-family, by incorporating time into their token causal definition. That is, BV−CM proposed in [BV18], distinguishes itself by extending causal models with a timing function. Thereby, allowing it to deal with a large range of controversial examples with time-critical scenarios. Moreover, rather than just building on the HP-definitions, this approach builds heavily on Hall's separation of causality into production and counterfactual dependence. From there they construct their definition according to a collection of necessary and sufficient conditions that are allegedly inherent to causation and are derived from common examples found in the literature.

Before discussing definitions that use languages other than causal models, it must be noted that there are additional, but rather rudimentary definitions of token causality. For reference see [HH11], [HH15], [Sch16] and [Wes15].

3.2.2 Definitions based on other Formal Languages

Among those token causality definitions that do not utilise causal models, there exists one block of formalisms using some kind of action language and another one that utilises languages from the CP-Logic family. Apart from that, there are two additional formal definitions and a multitude of (semi-)formal definitions.

Starting with the definitions that utilise action languages. In particular, there are two equivalent definitions, SC−ACC and SC−CF introduced in [BS18] and [KS20], that leverage the expressive capabilities of SC. Hence, both of them approach causality from a more procedural point of view. Their method for detecting token causes uses achievement causal chains, which can be roughly understood as a specially curated sequence of actions. Among the publications in set $\mathcal{F}$ this definition is one of the few that can identify causes expressed in first-order logic, as many of the other approaches are, as of now, bound to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

propositional level with no natural path for generalisation in sight, e.g. causal models. In [KS20] the authors introduce a revised definition that is not only built in the shadows of the counterfactual tradition but also capable of illustrating some of the accounts found in the regularity tradition. The approach presented in [LBV19], which shall be abbreviated with AT uses the action language $\mathcal{AL}$ and is therefore related to the situation calculus approaches. However, when contrasted against those, $\mathcal{AL}$ is, according to [LBV19], not only better equipped for representing indirect effects, but also commands a simple semantic.

Within the family of CP-Logic there are four definitions to be found. The first, called BV−11, one is formulated using CP and was defined in [Ven11]. Later the same authors presented a modified version of said definition in [BV12]. It is referenced here as BV−12. Those two definitions take an inherent probabilistic view on causation. Moreover, given the semantics of CP it can be argued that those definitions contain a procedural element as well. Akin to the definitions from Halpern and Pearl they later extended BV−12 by introducing normality, creating HH−CP in the process. Moving away from a probabilistic conception of causality, and more towards a process-orientated view. Another definition in the wider context of CP-Logic can be found in [DBV18, DBV19], their definition is according to its creators constructed with a regularity theoretical perspective in mind and is formulated using CP2. Since they call the semantic underpinning their approach the possible causal process semantics, their definition will be abbreviated with PCPS.

The definitions BCI formulated in [Boc18a] and BReg originating in [Bau13] rely on neither of the above mentioned modelling language families. Firstly, with BCI, Bochman coined a definition emerging out of the regularity theoretic tradition that uses two separate logics, namely causal theories introduced in [MT$^+$97] and his logic of causal rules introduced in [Boc04]. The latter language has close ties with the strongest input-output logic presented in [MvdT00b]. Bochman heavily emphasises that his approach is a regularity theoretic alternative to the counterfactual causal model approach. However, both the token causality definition, as well as the similarity to the input-output logic, suggest that his approach is closer tied to the counterfactual tradition that advertised. This is in stark contrast with BReg. Rather than relying on a specialised language with a complicated and heavy semantic machinery, Baumgartner took great effort in requiring only fairly common logical concepts. That is, his definition, which follows the regularity tradition, relies only on material implication and some minimality constraint, thus allowing him to construct his definition using only a slightly extended version first-order logic.

Before moving on to the more informal definitions, there are some that are difficult to place. Among the publications in the $\mathcal{F}$ those definitions are primarily discussed in [BV16]. All three of which were put forward by Hall, one termed Hall−07 was coined in [Hal07] and the other two, called here Hall−04p and Hall−04d, are taken from [Hal04]. Those definitions were originally defined using neuron diagrams and structural equations. The two latter definitions reflect Hall's view in [Hal04] that causality can be separated into different relations, namely production and dependence.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.2.3 Definitions based on Informal Languages

There are some definitions of token causality that are not completely formalised. Starting with the definition provided in [LSW19a], using the language FCA, it is categorised by its authors as semi-formal. Their definition called causal argument evaluation criteria or CAEC is strongly inspired by the NESS account. Moving fully into the informal realm, arguably the simplest definition in this category is the But-For or *sine qua non* test. Heavily used in the legal profession, it captures a highly simplified form of counterfactual reasoning. An improvement on the but-for test is the definition of Hart and Honore's called *causally relevant factors*, here abbreviated with CRF, which according to [WG17] has its origins in [HH59]. Later on, John Mackie introduced the so-called INUS-condition (Insufficient but Necessary part of an Unnecessary but Sufficient condition) in [Mac65]. Wright, inspired by both of those accounts, formulates his NESS-account (Necessary Element of a Sufficient Set) in [Wri87]. The latter two are both considered to be contributions to the regularity theoretic literature [Bau13].

Lastly the definition from [BVKPV10]. It exists at the intersection of logic and law and is formulated using CAR. The core idea is to check whether a given story (a chronological sequence of events) explains a designated set of propositions using the provided abductive causal theory. Then they use an abstract argumentation framework to rank the stories based on their compliance with evidence. It was not considered a token causality definition, as it is not intended for extracting causally significant events from the given story.

3.2.4 Categorisation

Here the definitions found in $\mathcal{F}$ are ranked based on their popularity. Moreover, they are categorised based on the languages used and based on their view on causality. For a quick overview consult Table 3.6, which summarised this subsection.

The popularity of a definition is determined by counting how many publications in $\mathcal{F}$ are mentioning the definition in question. To that extend Table 3.3 and Table 3.4 were constructed. Those tables depict which publication from $\mathcal{F}$ mentions which definition.

It is safe to say that HP-05, which is mentioned by 15 publications, is the most popular of the considered formalism by a considerable margin. The distant second place with 4 mentions is taken by an extension of HP-05, namely HP-05d. The third most popular definition is yet another one from Halpern and Pearl, i.e. HP-01. Making the dominance of Halpern's ideas throughout the field quite apparent. However, on closer look this assessment may change slightly. It is clear that this assessment favours old formalisms that were consistently discussed, investigated and refined. Since this could be (and most likely is) done by the same author one can recalculate this ranking while ignoring self-references. When corrected for this, i.e. when considering publications from other authors only, HP-05 is still referenced 11 times. However, the second place with 2 references is now shared by HP-15, Wood-03 and Hitch-01.

The definitions can also be differentiated based on the original language used to formulate

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

them. The data collected to do so can be viewed in Table 3.5. Here, it must be remarked that the depicted data are already aggregated based on language families. That is, rather than listing every variant in a particular language family, the family itself is used in the categorisation process. In particular, this affects definitions using one of the CP-Logics, and definitions using one of the many variants of causal models. A mere glance at the table suffices, to further strengthen the claim that the ideas put forward by Halpern and Pearl shape the literature around causality. Namely, a total of 16 definitions rely on causal models in one form or another. Removing all definitions formulated by either Halpern or Pearl, still provides us with a total of 12 definitions. This is in stark contrast with the second most popular modelling language family, CP-Logic. This language is only used by four definitions, all of which were either formulated by its creator. Among those languages with zero definitions under their belt, particular mention must be given to neuron diagrams. Firstly, although no definition found in $\mathcal{F}$ uses this language, it is often applied as a modelling tool for conveying the type causal relations of an example in an intuitive manner. Secondly, [EW10] formulated a definition for token causes using neuron diagrams, which was not captured by the outlined methodology.

Within the language families, the definitions can be distinguished further. In the CP-Logic family, `BV-11`, `BV-12` use the original formulation, `HH-CP` uses a slight extension of the original language that allows for the expression of norms and `PCPS` relies on the deterministic second language in this family. With respect to causal models the following differentiations can be made. `HP-05d`, `MbM` (and arguably `HP-15` as it can be extended to a definition that incorporates normality) rely on causal models extended by a normality ranking. `HmM` used causal models where variables have default values. Both `PAP` and `HP-05p` use some form of probabilistic causal models and `BV-CM` require their causal models to be extended by a timing function.

Finally, Table 3.6 provides a summary of the token causal definition discussed in this section. Additionally, it provides information about their age and provides a reference to the publication of origin (if available). Moreover, this table roughly categorises the discussed definitions based on their philosophical approaches, i.e. counterfactual, process orientated, probabilistic or regularity, in a rather naive and admittedly simplistic manner. That is, if a definition heavily relies on interventions or hypothetical statements, then it will be considered part counterfactual approach; if a definition views causal dependencies from a production point of view, i.e. an effect produced by a process triggered by its cause, or if it heavily emphasises time, then it will be considered part of the process-orientated approach; if a definition relies on probabilistic causal dependence relations to infer token causes, then it will be considered part of the probabilistic approach; if a definition uses some form of default reasoning, in particular if based on normality assertions, then it will be considered to be a member of the regularity approach. However, a statement by the author's of a definition declaring membership to one particular approach, is sufficient enough to classify a definition according to the authors assessment. As the approaches are (for the most part) not mutually exclusive, overlapping classification is possible. The purpose of this classification is to provide the reader with a quick intuition about the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

definition's capabilities and their tools for capturing token causality.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

	But-For	CRF	INUS	NESS	Hitch-01	HP-01	Wood-03	Hall-04p	Hall-04d	HP-05	Hall-07	HP-05d	Simple	SimpleJ
[VBD10]														
[BVKPV10]														
[LLLY10]														
[LY10]														
[GDG+10]							✓			✓			✓	✓
[CH10]														
[GL10]	✓									✓				
[HH11]	✓									✓		✓		
[Shu11]														
[Bri12]														
[Bau13]			✓	✓										
[HHEJ13]														
[HH15]										✓		✓		
[Wes15]					✓	✓	✓			✓				
[CFKL15]										✓				
[BV16]								✓	✓	✓	✓	✓		
[Sch16]														
[Hal16b]						✓				✓		✓		
[BS17]					✓									
[WG17]		✓	✓	✓										
[IKK17]														
[ACHI17]						✓				✓				
[FG17]										✓				
[LG17]	✓													
[Boc18a]				✓										
[II18]														
[BV18]										✓				
[Boc18b]				✓										
[DBV18]										✓				
[BS18]										✓				
[DBV19]														
[LSW19a]														
[LBV19]														
[LSW20]	✓	✓								✓				
[KS20]			✓							✓				
[II20]														

Table 3.3: Depicts which publication discuss which token causality definition (before 2011).

	PAP	BV-11	BV-12	BReg	PTC	HP-15	HH-CP	HP-05c	HP-05p	HmM	MbM	BV-CM	BCI	SC-ACC	PCPS	AT	SC-CF	CAEC
[VBD10]																		
[BVKPV10]																		
[LLLY10]																		
[LY10]																		
[GDG+10]																		
[CH10]																		
[GL10]																		
[HH11]																		
[Shu11]																		
[Bri12]																		
[Bau13]				✓														
[HHEJ13]																		
[HH15]																		
[Wes15]					✓													
[CFKL15]																		
[BV16]		✓					✓											
[Sch16]			✓															
[Hal16b]																		
[BS17]										✓								
[WG17]											✓							
[IKK17]																		
[ACH17]																		
[FG17]	✓								✓									
[LG17]								✓										
[Boc18a]													✓					
[II18]																		
[BV18]												✓						
[Boc18b]													✓					
[DBV18]														✓	✓			
[BS18]					✓													
[DBV19]														✓				
[LSW19a]															✓			
[LBV19]																✓		
[LSW20]																		
[KS20]					✓													
[II20]																	✓	✓

Table 3.4: Depicts which publication discuss which token causality definition (from 2011 onwards).

TU Bibliothek Your knowledge hub WIEN
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Articles	CM +*	CP/CP2	SC	CT	FOL+	ND	SM	AL	FCA	CAR
But-For										
CRF										
INUS										
NESS										
Hitch-01	✓									
HP-01	✓									
Wood-03	✓									
Hall-04p										
Hall-04d										
HP-05	✓									
Hall-07	✓									
HP-05d	✓									
Simple	✓									
SimpleJ	✓									
PAP	✓									
BV-11		✓								
BV-12		✓								
BReg					✓					
PTC	✓									
HP-15	✓									
HH-CP		✓								
HP-05c	✓									
HP-05p	✓									
HmM	✓									
MbM	✓									
BV-CM	✓									
BCI				✓						
SC-ACC			✓							
PCPS		✓								
AT								✓		
SC-CF			✓							
CAEC									✓	

Table 3.5: This table depicts which definitions rely on which language family

	Year	References	Language	Approach	Origin
But-For	?	4	NL	CF	?
CRF	1959	2	NL	RE	[HH59]
INUS	1965	3	NL	RE	[Mac65]
NESS	1987	4	NL	RE	[Wri87]
Hitch-01	2001	2	CM	CF	[Hit01]
HP-01	2001	3	CM	CF	[HP05]*
Wood-03	2003	2	CM	CF	[Woo05]°
Hall-04p	2004	1	?	CF	[Hal04]
Hall-04d	2004	1	?	CF	[Hal04]
HP-05	2005	15	CM	CF	[HP05]
Hall-07	2007	1	CM	CF	[Hal07]
HP-05d	2008	4	CM+N	CF,RE	[Hal08]
Simple	2010	1	CM	CF	[GDG$^+$10]
SimpleJ	2010	1	CM	CF	[GDG$^+$10]
PAP	2011	1	CM+P2	CF, PR	[TK11]
BV-11	2011	1	CP	CF, PR	[Ven11]
BV-12	2012	1	CP	CF, PR	[BV12]
BReg	2013	1	FOL+	RE	[Bau13]
PTC	2015	1	CM	CF,RE	[Wes15]
HP-15	2015	2	CM	CF	[Hal15a]
HH-CP	2016	1	CP+N	CF, PR, RE	[BV16]
HP-05c	2017	1	CM	CF	[FG17]
HP-05p	2017	1	CM+P	CF, PR	[FG17]
HmM	2017	1	CM+D	CF, RE	[BS17]
MbM	2017	1	CM+N2	CF, RE	[BS17]
BV-CM	2018	1	CM+T	CF, PO	[BV18]
BCI	2018	2	CT	RE (, CF)	[Boc18a]
SC-ACC	2018	1	SC	PO	[BS18]
PCPS	2018	2	CP2	PO, RE	[DBV18]
AT	2019	1	AT	PO	[LBV19]
SC-CF	2020	1	SC	CF, PO	[KS20]
CAEC	2020	2	FCA	RE	[LSW20]

Table 3.6: Summary of the token causality definitions discussed. The approaches considered are the Counterfactual approach (CF); a Process Oriented approach (PO); the Regularity Theoretic approach (RE); a (explicit) Probabilistic approach (PR). (*: Original in 2001; °: Original in 2003)

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.3 Token Causality: Benchmarks

The repeated attempts to formally capture token causality, not only produced a diversity of definitions and languages, but also a set of examples designed to benchmark the capabilities of token causal definitions. Those examples, increasingly complex, attempt to capture fragments of causality, as intuitively understood by humans. With many authors proposing new examples to highlight shortcomings of previously established formalisms, the literature concerning causation has amassed a wealth of such examples. Hence, this section shall provide an overview over some of the most prominent examples.

The first subsection will provide a quick overview of the Benchmarks highlighted in this section. Moreover, it will introduce a graphical language for encoding simple causal relations, called neuron diagrams. This language will be used in the subsequent subsections to highlight the structure of the causal relations within each example. The Sections 3.3.2, 3.3.3, 3.3.4, 3.3.5, 3.3.6, 3.3.7 & 3.3.8 all introduce scenarios that are commonly used to benchmark new token causality definitions. The intended purpose of those section is to provide an overview of the common examples and difficult edge cases found in the literature, as well as the discussion surrounding them. By contrast Section 4.2 will use the presented Benchmarks to compare a selection of token causality definitions. This section concludes by presenting some less prominent examples, which nevertheless highlight important aspects of causality.

3.3.1 Overview and Preliminaries

The literature is full of examples used to benchmark token causal definition, however, there is a small set of prominent examples, against which most definitions are tested. Those examples, capture aspects of causations, that seem to be relatively fundamental or particularly difficult to capture formally. In some cases, the difficulty can be attributed to the fact that there does not seem to be a consensus on what the correct intuitive answer to the questions raised by those examples might be. Some of those prominent examples are concerned with the following scenarios:

- *Symmetric Overdetermination*, which refers to the scenario where multiple processes, all of which producing the same outcome, terminate at the same time. It is discussed in Section 3.3.2 and will be represented by the Benchmark 3.3.1.
- *Switch*, which refers to the scenario where there exists an event that triggers one of two processes both of which have the same outcome, thus making the event immaterial for the outcome of the scenario. It is discussed in Section 3.3.3 and will be represented by the Benchmark 3.3.2.
- *Late Preemption*, which refers to the scenario where there are two causal processes running in parallel, both would produce the same outcome, but one process terminates before the other does. Thereby, bringing forth the outcome and rendering the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

second process irrelevant. It is discussed in Section 3.3.4 and will be represented by the Benchmark 3.3.3.

- *Early Preemption*, which refers to the scenario where there are two causal processes, both would produce the same outcome, but one process terminates before the other can even start. It is discussed in Section 3.3.5 and will be represented by the Benchmark 3.3.4.
- *Double Preemption*, which refers to the scenario where a process that would have prevented another process, was prevented by an entirely different process itself. It is discussed in Section 3.3.6 and will be represented by the Benchmark 3.3.5.
- *Bogus Preemption*, which refers to the scenario where an action is taken to interrupt an inactive process. It is discussed in Section 3.3.7 and will be represented by the Benchmark 3.3.6.
- *Short Circuit*, which refers to the scenario where an action is taken to prevent an inactive process, however, this triggers the process in the first place, which then has no effect because the original action prevents it from terminating. It is discussed in Section 3.3.8 and will be represented by the Benchmark 3.3.7.

Those Benchmarks were chosen based on how frequently they were discussed in the surveyed literature. Moreover, Table 3.7 provides an overview of which publication discusses which example. Hence, it not only allows one to gauge the popularity of each Benchmark, but also serves as a single point of reference for pointers into the literature.

Lastly, even among the less prominent examples, there are some that highlight questions that are important for formalising causality. For example, how to deal with two causal processes of different "strength"; can an omission be a cause; what constitutes as the same event; is there a size limit on a causally connected chain of events; does causality satisfy transitivity; is causality contrastive, i.e. given a situation, finding the cause of some event always requires another situation to contrast the original situation against; how do norms, normality and the attribution of guilt interplay with causality. Hence, such examples will be discussed briefly in Section 3.3.9.

Neuron Diagrams

As this chapter restricts itself to providing an overview of and the intuition behind the common examples in the literature, it is important to select a language that can convey the causal structure discussed in the examples in an intuitive manner. Given their simplicity and their graphical nature the language of neuron diagrams was selected for this task. Hence, the following introduces a slightly altered form of neuron diagrams found in the literature. However, the basic template is taken from [EW10].

A *neuron diagram* is similar to a labelled directed acyclic graph, where the vertices represent neurons. Every neuron is associated/labelled with a variable. An exogenous

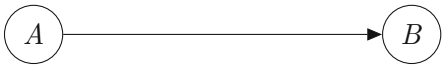
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

	Sym. Overdet.	Switch	Late Pre-emp.	Early Preemp.	Double Preemp.	Bogus Preemp.	Short Circuit
[VBD10]							
[BVKPV10]							
[LLLY10]							
[LY10]							
[GDG+10]		✓	✓		✓		
[CH10]		✓	✓		✓		
[GL10]							
[HH11]	✓	✓	✓			✓	
[Shu11]							
[Bri12]							
[Bau13]	✓	✓	✓	✓		✓	✓
[HHEJ13]							
[HH15]	✓			✓		✓	✓
[Wes15]	✓	✓	✓	✓		✓	✓
[CFKL15]	✓		✓			✓	
[BV16]			✓	✓			
[Sch16]							
[Hal16b]			✓				
[BS17]	✓			✓		✓	✓
[WG17]	✓		✓	✓			
[IKK17]	✓		✓	✓			
[ACHI17]			✓				
[FG17]				✓			
[LG17]							
[Boc18a]	✓	✓	✓	✓		✓	
[II18]							
[BV18]	✓	✓	✓	✓	✓	✓	✓
[Boc18b]			✓				
[DBV18]	✓	✓	✓		✓	✓	
[BS18]	✓	✓		✓			
[DBV19]	✓	✓	✓	✓	✓	✓	
[LSW19a]							
[LBV19]							
[LSW20]	✓		✓				
[KS20]			✓				
[II20]							

Table 3.7: This table summarises which paper in $\mathcal{F}$ discusses which example

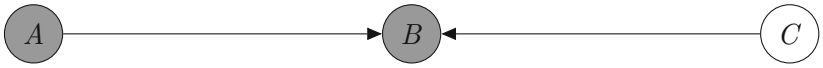
neuron is a neuron with no incoming edges (A), while an endogenous neuron must have at least one incoming edge (B).



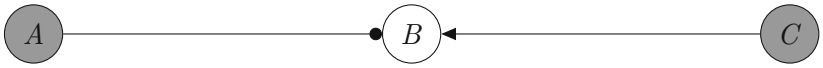
A neuron can be either active or not. If a neuron is active, it will be coloured grey (A). By default, all endogenous neurons are considered to be inactive (B). Whereas, the value of exogenous neurons is provided by the context of the situation.

Every endogenous neuron has a trigger threshold that indicates how many signals are required for the neuron to activate. An endogenous neuron with a single border requires a single signal (A), an endogenous neuron with a double border requires two signals (B), and an endogenous neuron with higher inertia has a double border and is annotated with a number (C).

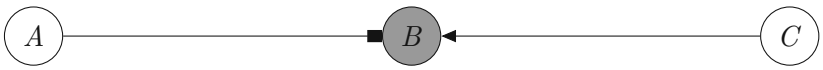
There are three kinds of relations. Firstly, stimulating edges, indicated by an arrow head, stimulate the target neuron, if the source neuron is active.



Secondly, inhibiting edges, indicated by a circle, prevent the target neuron to fire, if the source neuron is active.



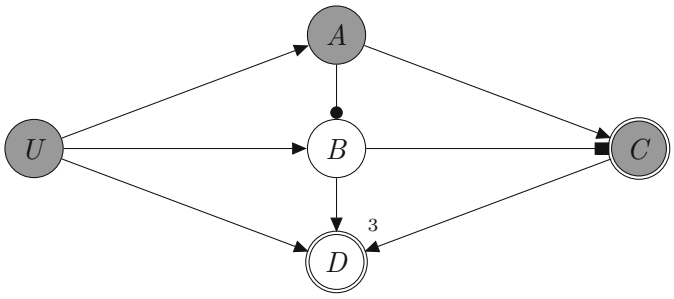
Thirdly, negating edges, indicated by a square, stimulate the target neuron, if the source neuron is not active.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

The most unorthodox choice made, was to add the negating edge to the language. While not necessary, it allows for cleaner modelling of some examples discussed in this section. To get accustomed to this informal definition consider Example 3.3.1.

Example 3.3.1. In neuron diagram depicted below, U is an exogenous neuron, while all others are endogenous. A and B fire on a single stimulus, C requires two and D needs 3 stimuli. Since U is active and is connected to A through a stimulating edge A receives a single stimulus, which in this case is sufficient for A to fire. Even though B received a stimulus from U it cannot be active, because it is connected to A via an inhibiting edge. C is active because it receives one stimulus from A and another from B, as it is connected via a stimulating edge to the former and a negating one to the latter. Finally, with U and C being the only stimulants for D the necessary threshold is not reached. Hence, D is not active.



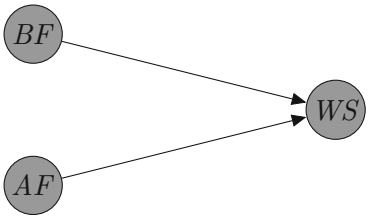
It may be of interest that [EW10] properly formalised (and generalised) neuron diagrams. To do so, they distinguished between neuron graphs capturing an abstract neuron structure and neuron diagrams representing the execution of neuron graphs for some set of inputs. This formalisation is of note, as it was used in [EW10] to create of yet another definition of token causality, which unfortunately was not captured by the outlined methodology (see Chapter 2).

3.3.2 Symmetric Overdetermination

Symmetric Overdetermination refers to the scenario, where multiple processes, all of which producing the same outcome, terminate at the same time. This is captured by the seemingly canonical situation presented in Benchmark 3.3.1, variants of which are discussed in [GDG+10, HH11, Bau13, HH15, Wes15, BS17, WG17, Boc18a, BV18, DBV18, BS18, DBV19, LSW20].

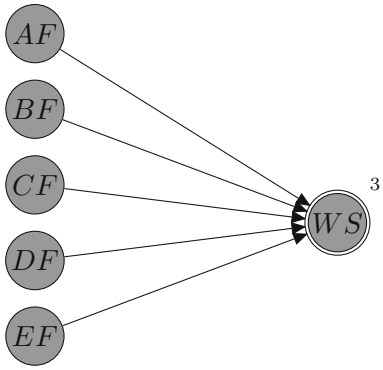
Benchmark 3.3.1. Alice (AF) and Bob (BF) each fire a bullet at a window, simultaneously striking the window, shattering it (WS). What caused the window to shatter?

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



This example is sometimes presented in an expanded form, which can be found in [GDG+10, CFKL15].

Example 3.3.2. Alice (AF), Bob (BF), Carol (CF), Dave (DF) and Eve (EF) all fire at a window. The window shatters after three hits (WS). What is the cause of the window shattering?



For scenarios of that kind it seems as if there does not exist a consensus on what a token cause should be [Hid05]. That is, it is unclear whether AF or BF individually should be considered a cause, whether the conjunct of AF and BF is the sole cause of WS or whether it is actually the disjunct that is the cause of WS.

The issue relates to a discussion about contributing causes, the intuition behind which is captured by the following story.

Example 3.3.3. Alice fills a sink with water. At each time interval Alice adds another drop of water. At one point the sink overflows. What caused the sink to overflow, was it only the last droplet or did the remaining droplets contribute to the outcome.

In Benchmark 3.3.1 if one allows for contributing causes then AF and BF individually could be considered as parts of the cause $AF \wedge BF$, but not causes themselves. Because intervening on a single variable does not prevent WS. Mirroring this [BV18] argue that while WS is not dependent on either AF and BF, both contribute to WS and thus both should be considered a contributing cause. Their definition, i.e. `PCPS` differentiates between counterfactually irrelevant and strongly counterfactually irrelevant variables. In

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

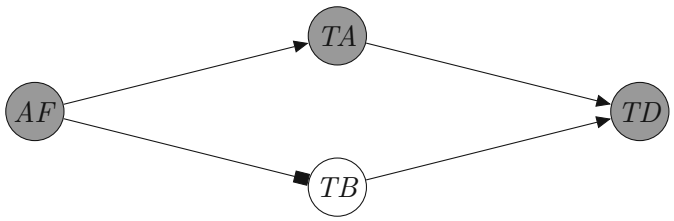
their view strongly counterfactually irrelevant variables should not be considered causes, while counterfactually irrelevant could still be considered causes. In this case, AF and BF are such counterfactually irrelevant variables.

To conclude, given the story presented in Benchmark 3.3.1 `HP-05`, `PTC`, `BV-CM`, `BCI` and `PCPS` claim that AF and BF individually are considered to be the cause of WS [BV18, Boc18a, DBV18, Wes15, Hal16a]. Moreover, Benchmark 3.3.1 reveals differences between Halpern's definitions. That is, `HP-05` considers AF and BF as the sole cause of WS, while `HP-15` only considers the conjunct $AF \wedge BF$ as the cause of WS. Moreover, both `HP-05` and `HP-15` consider $AF \vee BF$ to be a cause.

3.3.3 Switch

A Switch scenario seems to be characterised as follows. There is a variable representing some form of action, e.g. the flicking of a switch, irrespective of the variable's value a causal process is triggered. Each of those processes produce the same outcome. Hence, the original action was immaterial in the occurrence of said outcome. For the binary case, this effect can be observed in Benchmark 3.3.2, variants of which can be found in [GDG+10, HH11, Bau13, Wes15, Boc18a, BV18, DBV18, BS18, DBV19].

Benchmark 3.3.2. Alice flicks a switch (AF). The train travels on track A (TA), otherwise the train would have travelled on track B (TB). In both cases the train arrives at its destination (TD). Was AF the cause of TD?



With the flicking of the switch being immaterial, [BV18] postulates that most people would reject calling AF a cause of TD. Although this view is not uncontroversial, especially as embracing this intuition requires one to accept that causation is not transitive, i.e. it is clear that AF is the cause of TA, and TA is the cause for TD, yet AF would not be the cause of TD.

Moreover, the formalisation presented in Benchmark 3.3.2 is also subject of contention. That is, [HH11] suggest, that rephrasing the scenario in such a way that the variables TA and TB indicate whether the respective track is blocked or not, provides a more holistic model.

Example 3.3.4. Alice flicks a switch (AF). The train travels on track A (TA), otherwise the train would have travelled on track B (TB). Assuming that neither track$A$ (BA) nor track B (BB) are blocked, the train arrives at its destination (TD) in either of the two cases. Was AF the cause of TD?

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Another example of Switch which has an arguably less clear "solution", was discussed in [Wes15, Boc18a].

Example 3.3.5. Alice pushes Bob. Therefore, Bob is hit by a truck. Bob dies. Otherwise, Bob would have been hit by a bus, which would have killed him as well.

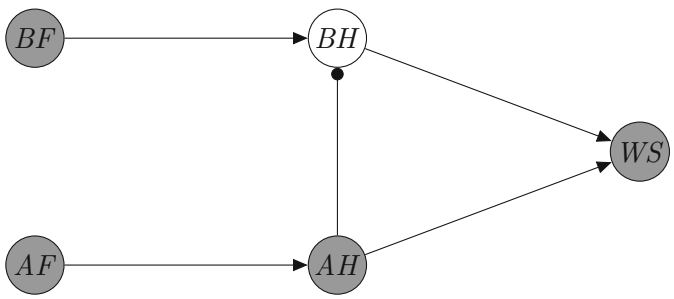
In Example 3.3.5 one is clearly faced with an instance of Switch. However, [McD95] claims that intuition would dictate that Alice did in fact kill Bob. [Wes15] argues that this intuition is a product of a hidden assumption, namely the hope there would be another option, like "Push Bob to safety". He therefore, claims that Example 3.3.5 is underspecified and would suggest declaring the available options as exhaustive within the model. That is, once the existence of any other unspecified option is excluded, AF should indeed be rejected as a cause. However, as long as there are other possibilities, AF could still be considered a cause.

To conclude, given the story of Benchmark 3.3.2, the definitions `HP-01` and `HP-05` claim AF to be the cause of TD. While the definitions `PTC`, `BV-CM`, `BCI`, `SC-ACC`, `SC-CF`, `HP-15` and `PCPS`, do not [BV18, Boc18a, DBV18, Wes15, Hal15a, BS18]. However, appeals to normality, can in some instances alleviate the deficits of `HP-01` and `HP-05`.

3.3.4 Late Preemption

Late Preemption, is quite similar to Symmetric Overdetermination, so much so that is sometimes referred to as Asymmetric Overdetermination [EW10]. They differentiate themselves, based on the fact that in Late Preemption the two running processes are not temporally aligned. That is, Late Preemption is a situation where two causal processes are running in parallel, both would produce the same outcome, but one process terminates before the other does. Thereby, bringing forth the outcome and rendering the second process irrelevant [BV18].

Benchmark 3.3.3. Alice (AF) and Bob (BF) each fire a bullet at a window. Alice's bullet hits the window first (AH). The window shatters (WS). Bob's bullet arrives second and does not hit the window (BH). What caused the window to shatter?

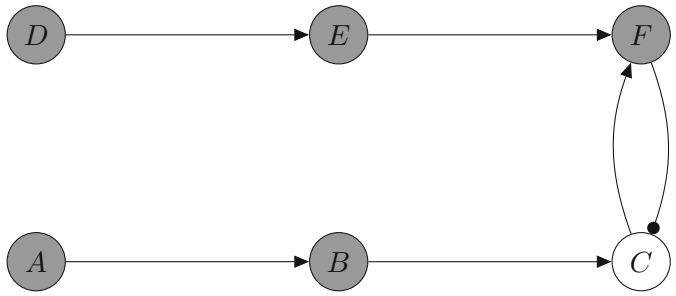


Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

There seems to be consensus on what the intuitive answer to the question should be. Namely, AF is the cause of WS. As clearly Alice's bullet prevents Bob's bullet to hit the window, by hitting it earlier. Hence, BF cannot be a cause of WS.

The addition of the variables AH and BH are vital, as their omission would produce an instance of Symmetric Overdetermination [HH11]. As observed by [BV18] the variable encoding Bob's failure of hitting the window, merely hides the fact that Bob was too late. That is, the addition of AH and BH simply hide the temporal aspect of the story, by implicitly encoding the order at which the bullets would hit the window, without explicitly engaging with time. In particular, [BV18] criticise this formalisation on the grounds that AF and BF trigger entirely different mechanisms. Hence, constructing a model that incorporates such a relationship is conceptually wrong. Favouring an alternative approach similar to the one suggested in [Hal16a, p. 34], where they encode temporal information into the model by introducing time-indexed variables.

In contrast to most other versions of Late Preemption, [Bau13] introduces a slightly different example of Late Preemption, which is included here for the sake of completeness. This example in particular is relatively unique, as it is one of the few that actually introduces a cyclic dependency between variables.



To conclude, the accounts `HP-05`, `HP-15`, `PTC`, `BV-CM`, `BCI`, `SC-ACC`, `SC-CF` and `PCPS` satisfy the provided intuition for the story presented in Benchmark 3.3.3. [BV18, Boc18a, DBV18, Wes15, KS20, Hal16a, p. 33]

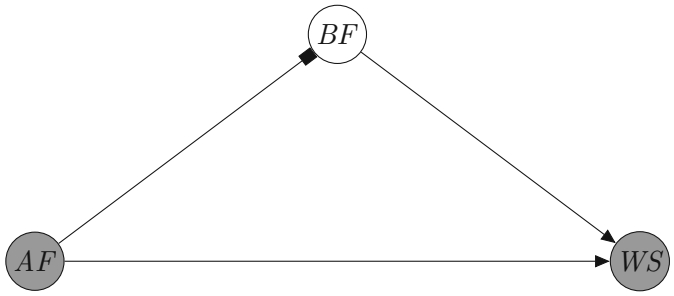
3.3.5 Early Preemption

Early Preemption refers to the scenario where there are two causal processes, both would produce the same outcome, but one process terminates before the other can even start. This is often captured by making the second process dependent on the first process. It is different to Late Preemption because the outcome of the two processes occurred before the second process was set in motion [BV18]. An alternative description, claims that the characteristic feature of Early Preemption is that the process in question is actually interrupted by another process, while in Late Preemption the process is never interrupted, it simply never has the opportunity to terminate [Bau13].

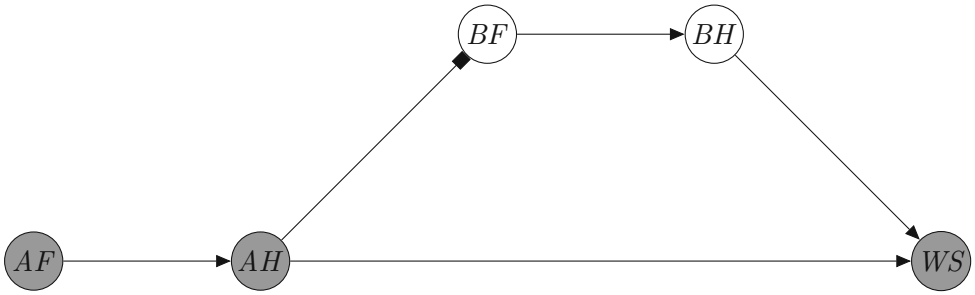
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Benchmark 3.3.4 seems to describe a canonical scenario for this effect, variants of it can be found in [Bau13, HH15, Wes15, BV16, BS17, WG17, FG17, Boc18a, BV18, BS18, DBV19].

Benchmark 3.3.4. (Early Preemption) Alice fires a bullet at the window (AF). If Alice hits the window (AH), the window shatters (WS). If Alice does not hit the window, Bob fires a bullet at the window (BF), hitting it (BH) leading to its shattering. What caused the window to shatter?



or a more complex version



Some authors consider Early and Late Preemption to be the same (or at least similar), thus they resolve examples discussing Early Preemption in a similar fashion. That is, AF is attributed to be the cause of WS, while BF is not considered to be a cause of WS. However, this straightforward analysis is deceptive. [BV18] noticed that Early Preemption has a close relationship with Switch. Assuming that Alice is certain that Bob will shoot at the window, if she neglects to do so, is faced with a choice. Either she shoots the window and it shatters or Bob will shoot at the window, shattering it in the process. Regardless, the status of the window is independent of her decision. In fact, Alice can only choose the causal path responsible for shattering the window, i.e. she can decide the how and not the if. Hence, it is a case of Switch. To further strengthen the similarity [DBV19] add an additional variable to the model, representing the bullet leaving the gun. In this case, let it be AH representing that Alice hit the window. This produces a model that is isomorphic to Benchmark 3.3.2. They claim, not uncontested,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

see [Wes15], that adding this variable should not influence the intuition about the causes at play.

Some try to appeal to probability in order to explain this discrepancy, i.e. they argue that it is implicitly assumed that causal process could fail, which subsequently pollutes the intuition [BV18, Hal07]. To contrast Benchmark 3.3.2 and 3.3.4, one could argue that people assume that Bob might fail to shatter the window, while the arrival of the train will always succeed. This view seems to be supported by the fact that if one attaches probabilities of arrival to the respective railway tracks found in Benchmark 3.3.2, some causal attribution to the switch event can be made. For example, if on track A the train has a 99% chance of arrival and on track B the train has a 1% chance, then Alice's flicking of the switch contributed to the train's arrival. How much the possibility of a process failing, influences human intuition can be observed in Example 3.3.6.

Example 3.3.6 ([BV18])**.** Suppose Alice reaches out and catches a passing cricket ball. The next thing on the ball's trajectory was a solid brick wall. Beyond that there was a window. Is Alice the cause of the window being intact?

People tend to classify this example as an instance of Switch. That is, catching the ball is immaterial for the status of the window. Alice merely decides the method of how the ball is stopped. However, by replacing the wall with another person Bob, this intuition shifts, declaring Alice's action to be causal for the well being of the window. The presumption is that this asymmetry arises due to the fact the prospect of the wall failing to stop the ball is not taken seriously [BV18, BS17]. Rather than relying on probabilities, this discrepancy could also be explained by somehow restricting the set of models under consideration, i.e. the fact that the wall fails to stop the ball, should not be considered as a possibility. A natural candidate for this task would be an appeal to normality.

Another method of dealing with this issue, is discussed in [DBV19]. They resolve this issue by distinguishing between enabling and triggering conditions. The former preempts the causal mechanism if not present, while the latter sets the mechanism in motion. Considering the discussed cases. Alice's flicking of the switch would be merely an enabling condition for the train taking track A or track B. By contrast, Alice's firing of the bullet is a triggering condition for the bullet hitting the window.

Example 3.3.7 further highlight the similarities and differences between Switch and Early Preemption.

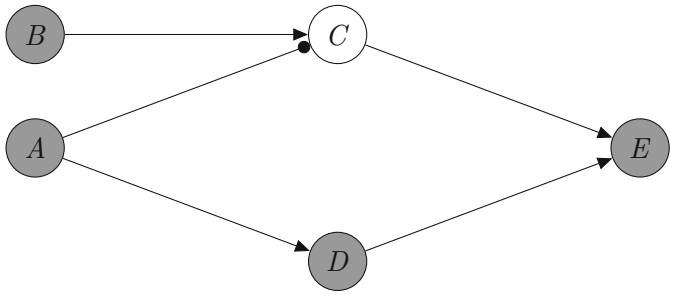
Example 3.3.7 ([Wes15])**.** Two two-state switches are wired to an electrode. The switches are controlled by A and B respectively, and the electrode is attached to C. A has the first option to flip her switch. B has the second option to flip her switch. The electrode is activated and shocks C if both switches are in the same position. B wants to shock C, and so flips her switch iff A does.

This example shares similarities with Switch and Early Preemption, it can be found in [Wes15, Boc18a]. While structurally similar to Early Preemption, one could argue that

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the action of A does not trigger the shocking of C and thus it should be considered to be a case of Switch. That is, A has no choice in the matter, and thus should not be considered a cause of C.

Furthermore, similar to Late Preemption, [Bau13] attributes Early Preemption a structure that is (slightly) different to the one presented in Benchmark 3.3.4.



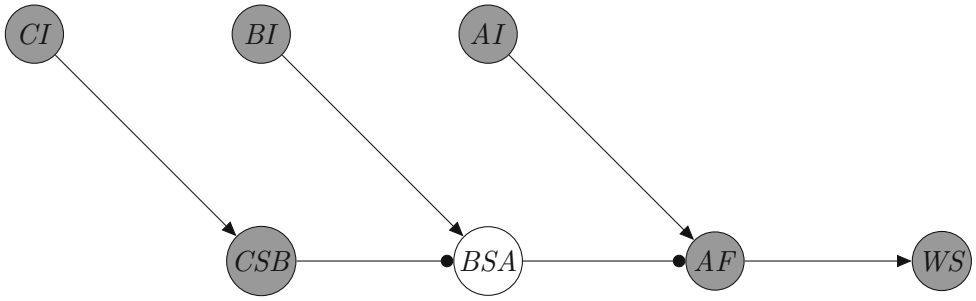
To conclude, the intuition presented in Benchmark 3.3.4 is satisfied by `HP-05`, `HP-15`, `PTC`, `BCI`, `SC-ACC`, `SC-CF` and `PCPS` [Boc18a, DBV18, Wes15, BS18]. This result is intentionally not shared by `BV-CM`. Because, as eluded to earlier, [BV18] understands the usual formalisation of this example as Switch. Hence, `BV-CM` does not declare AF to be the cause of WS. However, by properly extending the model with variables representing the accuracy of either Alice or Bob, AF becomes a cause. In the case of `BCI`, Alice is not only the cause of the window's shattering when she fires her bullet, but also when she does not.

3.3.6 Double Preemption

One speaks of Double Preemption if a process that would have prevented another process, was prevented by an entirely different process itself. Put differently, there are three active processes A, B, and C. Process C prevents process B from terminating and process B prevents process A from terminating. Since all three processes are active, process B is stopped from stopping process A, because process C is active, thus process A terminates unencumbered. That is, the potential preempter is preempted [DBV19]. Variants of Benchmark 3.3.5 can be found in [GDG+10, BV18, DBV18, DBV19].

Benchmark 3.3.5. Alice intends to fire a bullet at a window (AI). Bob intends to prevent Alice from hitting the window (BI). Bob tries to stop Alice (BSA). Bob is stopped by Carol (CSB). Alice fires a bullet (AF), hits the window (AH) and shatters it (WS). The window shatters (WS). What caused the window to shatter?

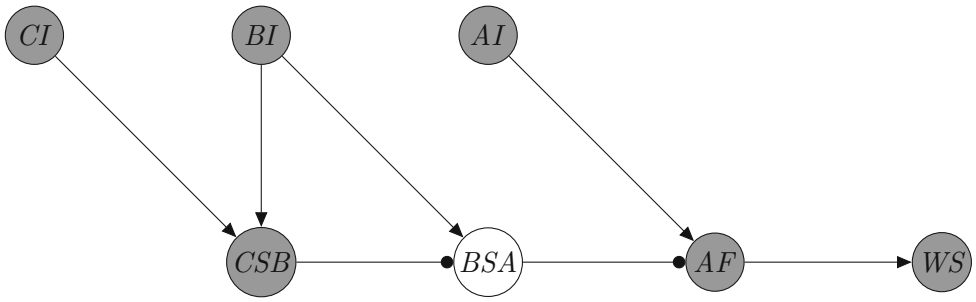
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



According to [Hal16a, p. 35], the intuition for this example is to attribute not only AI, but also CSB with being the causes of WS.

However, an issue arises in the case where Bob never intends to stop Alice, i.e. where BI never fires. Here intuition would dictate that CSB cannot be a cause of WS. This slight change of context produces counterintuitive inferences in some formalisms. Halpern suggests that this issue is the result of a too simplistic model. That is, Carol can only stop Bob, if Bob actually tries to stop Alice, a causal dependence is clearly not present in the model of Benchmark 3.3.5. A model including such a dependency can be seen in Example 3.3.8 [Hal16a, p. 36].

Example 3.3.8. Reformulation of Benchmark 3.3.5



[DBV18] extend the causal chain by adding a fourth party preventing Carol from stopping Bob, thereby creating Triple Preemption.

To conclude, in Benchmark 3.3.5 `HP-05` and `HP-15` deem AI and CSB to be the cause of WS. By contrast, `PCPS` does not consider CSB to be the cause of WS. However, they argue that their definition can easily be adapted into a state of compliance with Halpern's proposed intuition [DBV19, Hal16a, p. 36].

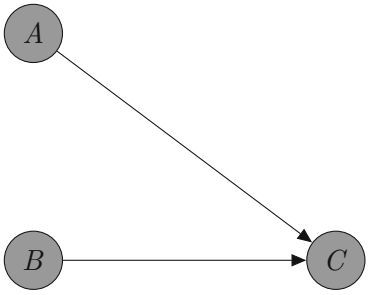
3.3.7 Bogus Preemption

Bogus Preemption occurs when an action is taken to interrupt an inactive process. Meaning the prevention is completely redundant, and therefore irrelevant for the outcome. Examples discussing Bogus Preemption can be found in [HH11, Bau13, HH15, Wes15,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CFKL15, BS17, Boc18a, BV18, DBV18, DBV19]. The canonical example, called "Careful Antidote", revolves around poisoned water.

Example 3.3.9. Alice is in possession of a lethal poison, but has a last-minute change of heart and refrains from putting it in Carol's water (A - A is true if Alice does *not* poison the water). Bob puts antidote in the water (B), which would have neutralized the poison. Carol drinks the water and survives (C - C is true if Carol survives).



The formalisation in Example 3.3.9 is used to demonstrate the limitation of structural equations, because in the structural equation framework this example is isomorphic to the example of Symmetric Overdetermination, while at the same time the intuition underlying both phenomena are vastly different. That is, Carol only dies if both Alice poisons the water and if Bob fails to add the antidote. Hence, C is only inactive, if both A and B are as well. Therefore, one is confronted with Symmetric Overdetermination scenario, indicating that A and B or their conjunct should be considered a cause of C. Yet, in the context of the given story, the supposed intuition dictates that neither A nor B should be considered a cause. [Bau13] elegantly observes that Symmetric Overdetermination discusses the overdetermination of occurrences, while Bogus Preemption is concerned with overdetermined absence [HH11, Wes15, HH15].

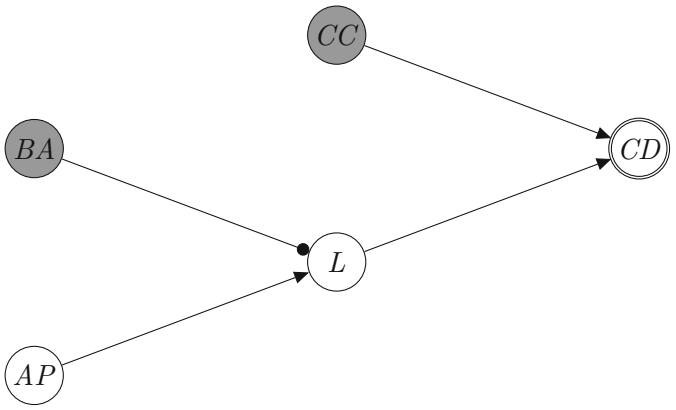
One suggested solution to this problem is to appeal to some notion of normality. That is, in addition to formalising the causal structure, it is necessary to provide the inference system with a theory of normality. The idea behind this approach is that one can use the notion of normality to exclude certain unreasonable contingencies. In this particular case, one could add a statement "Typically, people do not put poison in the water." to the model. Thus, the scenario where Alice actually poisons the water is less "normal" than the actual scenario. Hence, given this normality assumption, one can classify Bob's action as completely redundant. Thereby, excluding it from being a cause. [HH11, HH15].

Another suggestion to resolve this issue is to adapt the model used to represent the scenario. For example, the suitability of the presented model is directly criticised in [BS17], where it is called it impoverish. To rectify this [BS17] suggest the inclusion of a variable indicating the toxicity of the water. [HH15] notes, that such an extension is arguably more preferable than introducing normality. While not explicitly criticising the approach from Example 3.3.9, another frequent formalisation extends the model by a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

variable encoding the drinking of the water. The former is also found in [Boc18a, HH15], the latter is discussed in [DBV18, DBV19] and a combination of both can be found in [BV18]. Being the most detailed, a variant of the last is presented in Benchmark 3.3.6.

Benchmark 3.3.6. Alice intents to put lethal poison into Carol's water. However, Alice does not put lethal poison into Carol's water ($\neg AP$). Bob puts an antidote into Carol's water (BA). The water is lethal (L), if the poison is added without the addition of an antidote. If Carol would consumes the lethal water she would die (CD). Carol consumes her water (CC). Carol does not die ($\neg CD$).



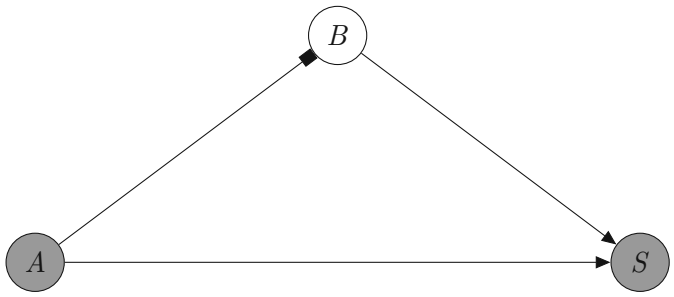
In [Wes15], they use the simplistic formalisation found in Example 3.3.9. Being isomorphic to Benchmark 3.3.1, their formalism, i.e. `PTC`, naturally concludes that both AP and BA causally influence the status of CD. In a formalisation obtained by extending the one found in Example 3.3.9 with a variable that holds when the poison is neutralised `HP-15` does not declare BA to be a cause. However, on the same model, both `HP-01` and `HP-05` consider AF to be a cause [Hal16a, p. 88]. In [Boc18a] they use yet again a slightly different formalisation. That is, utilising `CT` they construct a theory expressing that AP and not BA causes CD; not AP causes not CD; A and B cause not CD. Using this, `BCI` concludes that only the absence of Alice poisoning the water is the cause of Carol's survival. In [BV18], they use their timing function to differentiate between AP and BA. Therefore, if Alice's actions pre-date Bob's, then Alice's decision to refrain from poisoning the water is deemed to be the cause of Carol's survival by `BV-CM`. By contrast, if the order is reversed, the addition of the antidote would be classified as the cause. Additionally, if no timing is given this example is treated as a case of Symmetric Overdetermination. According to [DBV19], adding an antidote does interrupt the mechanism activated by poisoning the water. However, it is impossible for BA to preempt an inactive mechanism. Hence, only the refusal of Alice to poison the water should be considered a cause of Carol's survival. Their definition, i.e. `PCPS`, reflects this reasoning.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.3.8 Short Circuit

A short circuit scenario refers to a situation where an action is taken to prevent an inactive process. However, this triggers the process in the first place, which then has no effect because the original action prevents it from terminating. That is, the prevention creates its own relevance. Variants of this example, often called "Careful Poisoning", can be found in [Bau13, HH15, Wes15, BV18, BS17].

Example 3.3.10. (Careful Poisoning) Alice puts a harmless antidote in Carol's water (A). Bob intended to not put poison into the water. Seeing that the water contains an antidote Bob, adds the poison into the water (B - B holds if Bob does not administer the poison). This poison is countered by the antidote. Carol drinks the water and survives (S).

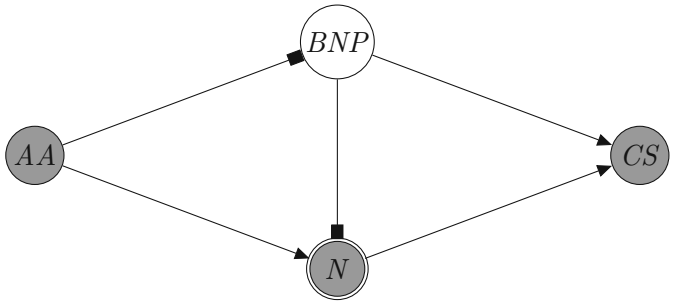


Similarly, to Bogus Preemption this formalisation of the presented story is isomorphic to the canonical Early Preemption case presented in Benchmark 3.3.4. This would suggest that adding the antidote to the water caused the survival of Carol. While not entirely uncontested, intuition would dictate that neither A nor B should be considered a cause of S [BV18].

The above instance of Short Circuiting is one of the examples referenced when talking about the limitations of structural equations and the necessity of extending causal models with some sense of normality ranking. However, [BS17] argue that the quality of the model is the source of the perceived similarities. They add another variable tracking the lethality of the water, i.e. is the water neutralised to resolve the issue of diverging intuitions on equivalent structures (see Example 3.3.11). That is, the original model fails to represent whether or not the antidote neutralizes the poison.

Example 3.3.11. Alice puts a harmless antidote in Carol's water (AA). Bob will not poison the water (BNP), if and only if Alice does not put the antidote into Carol's water The water will be neutralised (WN), if it contains both poison and antidote. Carol survives (CS), if either the water was neutralised or the water was not poisoned.

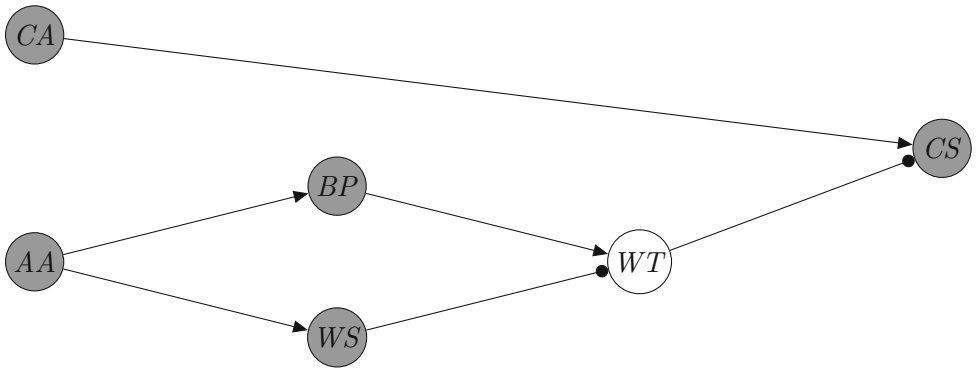
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



[BV18] takes a different angle. They argue that this issue has its origins in conflating Early Preemption and Switch. That is, if Alice adds the antidote, then Bob will add poison to the water, which promptly is neutralised allowing Carol to live on. Otherwise, Bob will not add poison to the water and Carol will be unscathed. Hence, the actions of Alice are immaterial to the well-being of Carol. Therefore, this example should be modelled as an instance of Switch, i.e. Alice merely decides in what way the water remains neutral. Thereby, realigning structure with intuition.

In [Bau13] another structure is classified under the umbrella of short circuit. Benchmark 3.3.7 labels the structure taken from [Bau13] to operate (roughly) within the narrative presented in Example 3.3.10.

Benchmark 3.3.7. Carol is alive (CA). Alice puts a harmless antidote in Carol's water (AA). Adding antidote to the water, protects it against poison (WS - "water save"). If Alice puts the antidote into Carol's water, Bob will poison the water (BP) Adding poison to an unprotected water makes it toxic (WT). If Carol would drink toxic water she would die (i.e. inhibiting CS). Carol consumes her water and survives (CS).



To conclude, given the formalisation found in Example 3.3.10 both `HP-05` and `HP-15` declare the addition of the antidote to be the cause of Carol's survival [Hal16a, p. 90]. The definition `PTC` arrives at the same conclusion [Wes15]. As already mentioned in [BV18], Example 3.3.10 is deemed to be a variant of Switch. Therefore, their formalism, i.e. `BV-CM`, is constructed in such a manner that adding the antidote is immaterial for Carol's survival.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3.3.9 Other Examples

The examples presented above occurred with the highest frequency in the surveyed literature. Naturally, there are also some examples that were discussed only sparingly. This subsection will serve as a quick overview of some interesting but relatively infrequent examples.

Starting with a simple example, originally given in the context of forest fires which is frequently used by Halpern, e.g. [HH11, HH15], as a benign introductory example.

Example 3.3.12. Alice (AF) and Bob (BF) each fire a bullet at a window, simultaneously striking the window. The window only shatters (WS), if it is hit by two bullets. What caused the window to shatter?

Already in such a small example, it becomes difficult to assess what a token cause should be. The first possibility would be to consider AF, BF and the conjunct of AF and BF as causes for WS. The second possibility is the rejection of the conjunct as cause for WS declaring only AF and BF as such. This can be motivated by the fact that if either AF or BF would not have occurred, then WS would have failed to happen as well. Hence, either Alice or Bob could have prevented the window from shattering, i.e. each action was essential for the outcome. The third possibility contrasts the previous one by declaring the conjunct as the sole cause of WS. The intuition behind this is that both actions are necessary for the window to shatter, i.e. both AF or BF must have been true to bring forth WS [Hal16a, p. 28]. Considering the last two possibilities. The first, essentially asks the question, what is necessary to prevent WS, whereas the second asks, what is necessary to bring forth WS.

The next example introduces the notion of trumping. Trumping is similar to Symmetric Overdetermination, but with the twist that the processes are now ranked. Namely, there are two processes active A and B, both processes produce the same outcome C. However, in the case that process A and process B conflict, the outcome of process A will always dominate.

Example 3.3.13. There are a left and a right window. Alice and Bob both order Carol to fire at the left window. Carol fires at the left window, shattering it. Commands from Alice always trump commands form Bob (e.g. if Bob would have ordered to fire at right window, Carol would still have fired at the left one.). Without a command Carol would not have fired at all. What caused the left window to shatter?

Example 3.3.13 is a reformulation of the canonical example found in [HH11, Wes15]. Here intuitions conflict whether one should consider Alice alone, both individually or Alice and Bob as a conjunction to be the cause of the left window shattering [HH11, Wes15].

Moreover, there is also disagreement on how trumping relates to other problem cases such as Symmetric Overdetermination and Preemption.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

This scenario presented in Example 3.3.13 clearly shares similarities with both Symmetric Overdetermination and Late Preemption. That is, similar to Late Preemption Bob's command starts a process that is "interrupted" by Alice's command. However, in this particular case the given commands are identical and are issued at the exact same moment, justifying the connection to Symmetric Overdetermination. [Hit11] argues that this example has far-reaching implications for the taxonomy of redundant causation. More precisely, he claims that this example cannot be classified as either Symmetric Overdetermination or Preemption. Hence, contradicting the belief that redundant causation is characterised by this dichotomy.

Particularly interesting is that in order to relegate the causal attribution in Example 3.3.13 to Alice and Bob alone, one must accept that Carol is void of agency. Essentially operating as a robot, with no responsibility being attributed to Carol. Such hidden assumptions, clearly tarnishes the intuitive interpretation of the example. Hence, it is critical to make such assumptions explicit. This highlights that using such stories, which are polluted with hidden assumptions, as a foundation for constructing causation may be problematic.

Causation by omission is the claim that the non-occurrence of an event caused another event, e.g. $\neg A$ causes B. Example 3.3.14 is commonly used to discuss whether omissions can be causes or not.

Example 3.3.14 ([HH15])**.** If there is hot weather, flowers will die. Watering prevents the flowers to die in hot weather. The neighbour does not water the flowers. The flowers die. What caused the flowers to die?

Is the demise of the flowers caused by the neighbour's neglect? While at first glance intuition would side for "yes", this question is not as straightforward as it may seem.

Not only does [BS17] claim that causation by omission is one of the open problems in determining actual causation. It is also the case, that there remains disagreement within the literature on whether causation by omission should be considered when defining token causality at all. [HH15] identified four established viewpoints within this debate. The first dismisses causation by omission, while the second completely embraces it. The third is positioned somewhere in between, declaring omissions to have some kind of secondary status. The last argues that it is the normative status of an omission that determines its causal status, e.g. in Example 3.3.14 the inaction of the neighbour only caused the death of the flowers, if he had the obligation to do so.

A person inclined to attribute the neighbour's omission as the cause of the flower's demise may have incorporated some implicit assumption into their reasoning process. That is, the sketched scenario never mentions that the neighbour is obliged to water the flowers, thereby placing the neighbour on equal footing with any other person on this world. Yet it seems natural to assume that, the neighbour was responsible to water the flowers. One suggestion, to ensure this implicit assumption is included, would be to appeal to

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

normality. That is, expecting the neighbour to water the flowers is considered less out of the ordinary than expecting some person on the other side of the world to do so. An additional benefit of this approach is that it provides sufficient flexibility to accommodate all the previously listed viewpoints. However, this flexibility can be problematic, as one has to rank scenarios based on their perceived normality. For example, what if the neighbour was a gardener, but also is sworn to kill this particular kind of flowers. Is the scenario now more or less normal? To avoid the reliance on normality, one could require that all variables used in the model are by default relevant for the scenario. Therefore, implying that the neighbour has some connection/obligation towards the flowers, by mere virtue of being considered in the model. Put differently, this would imply that only the neighbour and no other person is relevant for the status of the flowers [BS17].

Another example of omission is Example 3.3.15 which in some form or another can be found in [GDG+10, HH11, HH15, BS17].

Example 3.3.15 ([HH11])**.** Suppose that Billy is hospitalized with a mild illness on Monday; he is treated and recovers. In the obvious causal model, the doctor's treatment is a cause of Billy's recovery. Moreover, if the doctor does not treat Billy on Monday, then the doctor's omission to treat Billy is a cause of Billy's being sick on Tuesday. But now suppose that there are 100 doctors in the hospital. Although only doctor 1 is assigned to Billy (and he forgot to give medication), in principle, any of the other 99 doctors could have given Billy his medication. Is the nontreatment by doctors 2–100 also a cause of Billy's being sick on Tuesday?

Not only does Example 3.3.15 demonstrate the issues with omission, but it also raises questions about modelling, i.e. what should be included in our models and what does it mean for something to not be included in the model. Moreover, it further demonstrates how norms can influence our causal intuition, i.e. because Billy is assigned a specific doctor, all other doctors are suddenly void of the obligation to help Billy in this regard.

Another example, highlighting norms and expectations and their influence over causal attribution is Example 3.3.16, which can be found in [BV16, HH15].

Example 3.3.16 ([HH15])**.** The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly informed them that only administrative assistants are allowed to take the pens. On Monday morning, one of the administrative assistants encounters professor Smith walking past the receptionist's desk. Both take pens. Later, that day, the receptionist needs to take an important message...but she has a problem. There are no pens left on her desk.

In this example, intuition would dictate that professor Smith caused the absence of pens. In [KF08] it was empirically tested and confirmed that this seems to be the judgement most humans would make.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

An entirely different area of discussion is summarised by Example 3.3.17. It represents the stream of discussions relating to whether and how one should distinguish between causes and background conditions.

Example 3.3.17 ([HH15])**.** Consider a fire that is caused by a lit match. While the fire would not have occurred without the presence of oxygen in the atmosphere, the oxygen is deemed to be a background condition, rather than a cause.

Depending on the position taken in this discussion, one would either accept or deny the presence of oxygen the status of cause [HH15]. Obviously, this relates to the ideas of causes and contributing causes. To be more precise, while the oxygen contributed to the fire, it was merely a static precondition, and therefore insufficient of explaining a newly occurring event, i.e. one cannot explain change with a constant.

Speaking of change Example 3.3.18, taken from [HH11], stirs one directly into a metaphysical discussion about what an event is. Meaning, if the occurrence of an event is delayed (even by a tiny amount), is it still the same event. Neglecting this distinction during the modelling process may result in undesirable results.

Example 3.3.18. Alice plans to go camping in June (AC). If there is a forest fire in May (FF_m), Alice will not go camping. If Alice goes camping, she will cause a forest fire (FF_j).

That is, if one would not have explicitly distinguished between the forest fire in May and the forest fire in June by using separate variables, the model would contain circularities and thereby allows for counter-intuitive inferences, such as creating a forest fire in June causes Alice to go camping [HH11].

Moreover, this further raises the question of whether there can be a restriction on the distance between cause and effect, e.g. temporal distance. Example 3.3.19 represents a case where the "cause" is so far removed from the effect, that it is questionable to call that event a cause in the first place.

Example 3.3.19 ([HH15])**.** A lit match aboard a ship caused a cask of rum to ignite, causing the ship to burn, which resulted in a large financial loss by Lloyd's insurance, leading to the suicide of a financially ruined insurance executive. The executive's widow sued for compensation, and it was ruled that the negligent lighting of the match was not a cause (in the legally relevant sense) of his death.

The answer to whether the sailor dropping a lit match should be charged with being the cause of another person's death is yet again uncertain. Due to its long causal chain, the cause is so far removed from the effect that intuition would disagree with declaring the sailor to be a cause. That is, does the causal signal decrease while moving along a causal chain. This discussion shares some similarities with Minsky's account of "commonsense" reasoning in [Min07], where he argues that in contrast to logical reasoning, arguments

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

performed using "commonsense" reasoning gradually lose their viability as the argument chain increases in size. Hence, to ensure that the chain remains intact, additional arguments, are required to support the whole structure. A similar interpretation may be apt for causal reasoning as well.

However, another difference to conventional logical reasoning is the issue of transitivity. As already alluded to in the discussion of Benchmark 3.3.2, causation may not be transitive. Example 3.3.20 is another scenario that demonstrates that causation seems to violate transitivity [Hit01]. However, this example has a structure different to Switch and seems to share a close similarity with short-circuiting.

Example 3.3.20 ([GDG+10])**.** A boulder slides toward a hiker, who, seeing it, ducks. The boulder misses him and he survives. Did the boulder sliding cause his survival?

Clearly, intuition would dictate that the boulder is not the cause of the hiker survival. Yet the boulder caused the hiker to duck and ducking ensured that the hiker survived. Furthermore, note that this introduces a rather interesting pattern. Namely, the boulder initiates a process, and active intervention is required to prevent this process from terminating. That is, we have here a scenario, where a process triggers a process that prevents its termination, i.e. short-circuiting.

In [BS17] Example 3.3.21 was used to argue that causation should be considered as contrastive. Meaning it is not a binary, but a tertiary relation, i.e. to identify causes in a given situation, requires one to contrast the actual situation against another one. Hence, Example 3.3.21 is designed to induce two different intuitions depending on which possible scenario is used to counterfactually contrast the actual scenario against [BS17].

Example 3.3.21 ([BS17])**.** Consider a case where doctor can administer no dose, one dose, or two doses of medicine to patient. patient will fail to recover if no dose is administered, but will recover if either one or two doses are administered. Let us suppose that doctor in fact administers two doses and Patient recovers.

Here in particular, giving the patient two doses rather than zero doses caused the patient to recover. However, administering two doses rather than one dose did not cause the patient to recover [BS17].

The last example, Example 3.3.22, is relatively common, being discussed in [Wes15, CFKL15, Hal16b, Boc18a]. It is of particular relevance, as it was this example that lead to a reformulation of `HP-01`. While `HP-01` struggled with this example, its successor `HP-05` was able to match the intuitive answer. However, [Hal16b] demonstrated that with proper modelling `HP-01` can achieve the same inferences as `HP-05` on this example.

Example 3.3.22 ([Wes15])**.** A firing squad consists of shooters B and C. It is A's job to load B's gun, C loads and fires his own gun. On a given day, A loads B's gun. When the time comes, only C shoots the prisoner.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Here intuition would dictate that C was the one causing the prisoners death.

This chapter provided a wide overview of the different languages, definitions and benchmarks found in the causality literature. Moreover, the categorisations provided in this chapter allows us to intelligently select a few such constructs for a more detailed and technical investigation, which will be carried out in Chapter 4.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 4

Comparing recent systems for token causality

This chapter describes a selection of token causal definitions and their corresponding languages. To that end, Section 4.1 provides the needed background required to understand how those definitions perform causal inference. Section 4.2 compares the introduced formalisms by applying each of them to the benchmarks introduced in Section 3.3.

In order to test the behaviour of a definition on a particular benchmark, we are required to model the described scenario using the respective languages. However, as discussed in Section 3.3, even this task is subject of contention. Hence, we tried to ensure that the respective formalisations are similar in structure. This required us to sometimes diverge from the formalisations found in the literature, assuming they were discussed in the first place (see Table 3.7). By following this approach, we could observe that the discussed definitions exhibited similar behaviour across many benchmarks. This is particularly true for the older benchmarks, e.g. Symmetric Overdetermination, Late and Early Preemption. By contrast, we could observe a slightly more diverse set of results on the newer benchmarks, e.g. Bogus Preemption and Short Circuit.

4.1 Definitions

We discuss the token causality notions from each of the three most popular language families (see Chapter 3): the one developed in [BV18] which relies on causal models; the one introduced in [DBV19] which uses CP-Logic; the one discussed in [Boc18a] which builds on Non-Monotonic Theory. Moreover, due to the influence of its successors we also discuss the newest theory put forward by Halpern, which was introduced in [Hal15a].

This section unfolds as follows. As each formalism was introduced using a slightly different notation, first Section 4.1.1 provides a homogenised notation to ensure that

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

similar concepts across the various formalisms are expressed in a similar manner. Causal models (CM) and the modified Halpern and Pearl definition for token causality (HP-15) are presented in Section 4.1.2. Building on the definition of causal models, Section 4.1.3 will discuss BV-CM, the most recent token causal definition using CM, which was introduced in [BV18]. Section 4.1.4 presents the latest entry in the CP-Logic family (CP2) and its corresponding token causality definition PCPS, developed in [DBV19]. Lastly, Bochman's definition for causal inference (BCI) taken from [Boc18a], and its corresponding language, non-monotonic causal theories (CT), is introduced in Section 4.1.5.

4.1.1 Notation

By convention, sets will be written using capital letters and are usually referred to using the symbols X, Y and Z. Variables are written in lower case and are commonly referenced using the letters u, v, x, y and z. Values of variables are written in lower case and bold, e.g. given the variable x, $\mathbf{x}$ is the value of the variable x. In the binary case, the variable x can either have the value **true** or **false**. Implicitly, **true** will be treated as 1 and **false** as 0.

Exogenous variables are variables whose values are determined by factors outside of the model. Whereas endogenous variables are determined by the values of exogenous variables using the rules of the modelled system.

Definition 4.1.1. A signature $\Sigma := (\mathcal{U}, \mathcal{V}, \mathcal{R})$ consists of set of exogenous variables $\mathcal{U}$ and a set of endogenous variables $\mathcal{V}$. Let $\mathcal{W} := \mathcal{U} \cup \mathcal{V}$. Moreover, $\mathcal{R}$ is a function that specifies the range of each variable in the signature Σ, i.e. $\forall x \in \mathcal{W}\ \mathcal{R}(x) = X$ for some non-empty set X.

There are two additional notion of a signature used in this thesis.

Definition 4.1.2. A binary signature $\Sigma := (\mathcal{U}, \mathcal{V})$ is a signature where each variable is binary, i.e. $\forall x \in \mathcal{W}\ \mathcal{R}(x) = \mathbb{B}$. A propositional signature $\Sigma := \mathcal{W}$ is a binary signature without an endogenous-exogenous variable distinction.

Sometimes it will be assumed that the variables in $\mathcal{W}$ adhere to some implicit total order, i.e. the set $\mathcal{W}$ can equally be understood as the tuple $\vec{w} = (x_1, \dots x_{|\mathcal{W}|})$ containing the elements in $\mathcal{W}$. In a slight abuse of notation this distinction is not made explicit. For example, for some $X \subseteq \mathcal{W}$ and some function f, a statement such at $\bigtimes_{x \in X} f(x)$ is essentially $\bigtimes_{i=1 \wedge x_i \in X}^{|\mathcal{W}|} f(x_i)$.

For some signature Σ, let σ be a function assigning a variable some value, i.e. for $x \in X \subseteq \mathcal{W}\ \sigma(x) \in \mathcal{R}(x)$. If σ is only defined over $\mathcal{U}$, then it will be called context. If it is defined for all variables in $\mathcal{W}$ it will be called an assignment. Moreover, the encoding of an assignment as a set of literals (in the binary case) is referred to as world, which in a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

slight abuse of notation will also be indicated by σ. That is, σ is a world, if for all $x \in \mathcal{W}$ either $x \in \sigma$ or $\neg x \in \sigma$, but not both.

The symbol $\mathcal{L}$ is used to a indicate formal language. Languages are constructed overs signatures and contain formulas, which are indicated using φ, ϕ and χ. As an example of such a formal language consider $\mathcal{L}_{\mathbb{B}}$, the language of classical propositional logic.

Definition 4.1.3. The language $\mathcal{L}_{\mathbb{B}}$ can be recursively constructed over a propositional signature Σ. That is,

- $\bot \in \mathcal{L}_{\mathbb{B}}$, $\top \in \mathcal{L}_{\mathbb{B}}$ and $\Sigma \subseteq \mathcal{L}_{\mathbb{B}}$
- if $\varphi \in \mathcal{L}_{\mathbb{B}}$ then $\neg\varphi \in \mathcal{L}_{\mathbb{B}}$.
- if $\varphi, \psi \in \mathcal{L}_{\mathbb{B}}$, then for $\circ \in \{\wedge, \vee, \rightarrow\}$ $\varphi \circ \psi \in \mathcal{L}_{\mathbb{B}}$.

The term literal in a binary variable context, references either the variable itself or the negated variable, i.e. for some variable x the literal l is either a positive literal $l = x$ or a negative literal $l = \neg x$. In this thesis literals are commonly indicated by the letters l, p and q, while a set of literals is denoted using L.

The structures used to interpret such languages vary greatly. However to indicate such structures the symbol $\mathcal{I}$ is used. In the case of $\mathcal{L}_{\mathbb{B}}$, $\mathcal{I} = \sigma$. Moreover, for the sake of completeness consider the following definition of the semantics of propositional logic.

Definition 4.1.4. A sentence $\varphi \in \mathcal{L}_{\mathbb{B}}$ is evaluated under an interpretation $\mathcal{I}$ such that

- if $\varphi = \bot$ then $\mathcal{I}(\bot) =$ **false** and if $\varphi = \top$ then $\mathcal{I}(\top) =$ **true**
- if $\varphi = \neg\psi$ then $\mathcal{I}(\neg\psi) =$ **true** $\iff \mathcal{I}(\psi) =$ **false**
- if $\varphi = \psi \wedge \chi$ then $\mathcal{I}(\psi \wedge \chi) = \min\big(\mathcal{I}(\psi), \mathcal{I}(\chi)\big)$
- if $\varphi = \psi \vee \chi$ then $\mathcal{I}(\psi \vee \chi) = \max\big(\mathcal{I}(\psi), \mathcal{I}(\chi)\big)$
- if $\varphi = \psi \rightarrow \chi$ then $\mathcal{I}(\psi \rightarrow \chi) =$ **false** $\iff \mathcal{I}(\psi) =$ **true** and $\mathcal{I}(\chi) =$ **false**

Lastly, the symbol Δ will be used to indicate a set encoding causal relationships, irrespective of whether it is provided by the model, e.g. causal models, or by a set of formulas, e.g. Bochman's causal inference. Moreover, for some (endogenous) variable x, let δ_x be the set of causal rules influencing x. If $|\delta_x| = 1$ then let δ_x be the rule itself.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.1.2 The modified Halpern and Pearl definition

Here we present HP-15 and CM, unless otherwise specified the definitions are taken from [Hal15b].

A causal model can be viewed as a tuple consisting of a set of exogenous variables $\mathcal{U}$ and endogenous variables $\mathcal{V}$, where each variable is assigned a range of possible values by the function $\mathcal{R}$. The former are variables whose values are specified using the context of the modelled situation. The latter are variables whose values are determined by exactly one structural equation from the set of structural equations Δ. Such equations are best understood as assignments, determining the value of an endogenous variable based on the values of all other variables in the model.

Definition 4.1.5. A causal model $\mathcal{I} := (\Sigma, \Delta)$ is a pair, where $\Sigma := (\mathcal{U}, \mathcal{V}, \mathcal{R})$ is a signature and Δ is a set of modifiable structural equations containing for each variable $x \in \mathcal{V}$ a single function $\delta_x : \times_{y \in \mathcal{W} \setminus \{x\}} \mathcal{R}(y) \to \mathcal{R}(x)$.

Usually, the function δ_x will be expressed using a shorthand notation. That is, for the causal model $\mathcal{I}$ where $\mathcal{U} := \{u\}$ and $\mathcal{V} := \{x, y, y'\}$, the function $\delta_x(y, y', u) = y + u$ is abbreviated as $x := y + u$. To emphasise that this structural equation is not an equation in the classical algebraic sense the equation is written as $x := y + u$, an idea taken from [Wes15]. In its most general form, a causal model could look as follows.

Example 4.1.1. A simple causal model capturing the fact that firing a bullet at a window (BF) causes it to shatter (WS) could take the form $\mathcal{I} := (\Sigma, \Delta)$ with $\Sigma := (\{x_{BF}\}, \{x_{WS}\}, \mathcal{R})$ such that $\mathcal{R}(x_{BF}) = \mathcal{R}(x_{WS}) := \{\textbf{true}, \textbf{false}\}$ and with $\Delta := \{x_{WS} := x_{BF}\}$. However, in general we are not restricted to such scenarios. For example, if we would like to model the relationship between altitude (A) and temperature (T) in a linear manner, we could choose to do this with the model $\mathcal{I} := (\Sigma, \Delta)$ with $\Sigma := (\{x_A\}, \{x_T\}, \mathcal{R})$ and $\mathcal{R}(x_A) = \mathbb{R}$ and $\mathcal{R}(x_T) = \mathbb{R}^+$ and with $\Delta := \{x_T := \gamma \cdot x_A\}$ where γ is some constant.

When distinguishing between endogenous and exogenous variables, Halpern takes a rather pragmatic approach and uses a single exogenous variable. The range of this variable is chosen such that it can encode all possible variable-value combinations. For example, to encode a set of X variables into a single variable, Halpern compresses those into a tuple $\vec{x} \in \times_{x \in X} \mathcal{R}(x)$

Emerging out of the counterfactual tradition, it is natural that the language provides a convenient method to model interventions. Those interventions are akin to asking questions such as, how the model would change if a certain variable is fixed to a particular value.

Definition 4.1.6. Let $\mathcal{I} := (\Sigma, \Delta)$ be a causal model, then the model $\mathcal{I}_{x:=\mathbf{x}} := (\Sigma, \Delta_{x:=\mathbf{x}})$ where $\Delta_{x:=\mathbf{x}}$ is identical to Δ, but for the structural equation of x which is fixed to $x := \mathbf{x}$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Multiple interventions at the same time, i.e. $\mathcal{I}_{x_1:=\mathbf{x}_1,\dots,x_k:=\mathbf{x}_k}$ for some $k \in \mathbb{N}$, will be written as $\mathcal{I}_{x_1:=\mathbf{x}_1,\dots,x_k:=\mathbf{x}_k}$ or if clear from the context it will be abbreviated using vector notation $\mathcal{I}_{\vec{x}:=\vec{\mathbf{x}}}$. Δ is treated in analogously.

Example 4.1.2. Consider the causal model of a simple scenario described in [BV18]. This model contains two binary exogenous variables x_{AF} and x_{BF} and one binary endogenous variable x_{WS}. The value of the endogenous variable is characterised using the structural equation $x_{WS} := x_{AF}$. That is, we have the following causal model $\mathcal{I} := (\Sigma, \Delta)$ with $\Sigma := (\{x_{AF}, x_{BF}\}, \{x_{WS}\}, \mathcal{R})$ and $\mathcal{R}(x_{AF}) = \mathcal{R}(x_{BF}) = \mathcal{R}(x_{WS}) = \{\mathbf{true}, \mathbf{false}\}$ and with $\Delta := \{x_{WS} := x_{AF}\}$. By intervening on x_{AF} and fixing it to **false** we modify the equation of x_{AF} which results in the model $\mathcal{I}_{x_{AF}:=\mathbf{false}} = (\Sigma, \Delta_{x_{AF}:=\mathbf{false}})$ where $\Delta_{x_{AF}:=\mathbf{false}} = \{X_{AF} := \mathbf{false}\}$.

Computing the values of the variables present in a causal model requires context, i.e. an assignment of values to all exogenous variables. Once such an assignment is given, the values of the endogenous variables are derived by finding a solution for the set of structural equations.

Definition 4.1.7. Given a causal model $\mathcal{I} := (\Sigma, \Delta)$. Let σ be a setting of the exogenous variables in $\mathcal{I}$ such that for $x \in \mathcal{U}$ $\sigma(x) \in \mathcal{R}(x)$, which is referred to as context. Then $(\mathcal{I}, \sigma)$ is a causal model with context.

Example 4.1.3. Consider the simple window model $\mathcal{I} := (\Sigma, \Delta)$ from Example 4.1.2. A suitable context would be $\sigma := \{x_{AF} \mapsto \mathbf{true}, x_{BF} \mapsto \mathbf{false}\}$.

It is in general possible to have cyclic interdependences between variables, the most discussed subset of causal models has a clear dependence hierarchy that prevents such cyclic relations and ensures that the values of the variables are uniquely determined by the context.

Definition 4.1.8. A causal model $\mathcal{I}$ is acyclic if there is some total ordering $\prec$ of the endogenous variables such that if $x \prec y$, then x is independent of y, i.e. $\forall \mathbf{y}, \mathbf{y}' \in \mathcal{R}(y)$ $\delta_x(\dots, \mathbf{y}, \dots) = \delta_x(\dots, \mathbf{y}', \dots)$.

An alternative definition can be found in Halpern's book [Hal16a] where he defines recursive and strongly recursive models, with the latter being a special case of the former.

Example 4.1.4. The causal model from Example 4.1.2 is acyclic. By contrast, the model induced by the set of structural equations $\Delta := \{B := \gamma \cdot C + A, C := \eta \cdot B\}$ would be considered cyclic.

Although most definitions remain the same for cyclic models, *unless otherwise specified, any subsequent reference of causal models refers to acyclic causal models.*

Causal models are only concerned with encoding type causal relations among variables. Hence, in order to express and evaluate causal claims over those structures, including

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

interventions, an appropriate language is required. This language is constructed over a signature and is essentially an extension of classical propositional logic, allowing one to query the values of variables and enabling one to express interventions.

Definition 4.1.9. Let $\Sigma := (\mathcal{U}, \mathcal{V}, \mathcal{R})$ then language $\mathcal{L}_{\text{CM}}$ can be constructed as follows. Let $\mathcal{L}_{\text{cm}} \subseteq \mathcal{L}_{\text{CM}}$ be recursively defined as:

- $x = \mathbf{x} \in \mathcal{L}_{\text{cm}}$, with $x \in \mathcal{V}$ and $\mathbf{x} \in \mathcal{R}(x)$ are referred here as literals (or primitive events);
- if $\varphi \in \mathcal{L}_{\text{cm}}$ then $\neg\varphi \in \mathcal{L}_{\text{cm}}$;
- if $\varphi, \psi \in \mathcal{L}_{\text{cm}}$, then for $\circ \in \{\wedge, \vee, \rightarrow\}$ $\varphi \circ \psi \in \mathcal{L}_{\text{cm}}$.

$\mathcal{L}_{\text{CM}}$ extends $\mathcal{L}_{\text{cm}}$ as follows. For all $\varphi \in \mathcal{L}_{\text{cm}}$, one has

- $\varphi \in \mathcal{L}_{\text{CM}}$ and
- $[y_1 := \mathbf{y}_1 \wedge \cdots \wedge y_k := \mathbf{y}_k]\varphi \in \mathcal{L}_{\text{CM}}$

and with $y_1 := \mathbf{y}_1, \ldots, y_k := \mathbf{y}_k$ being distinct variables in $\mathcal{V}$ and $\mathbf{y}_i \in \mathcal{R}(y_i)$ for $1 \leq i \leq k$. Such a causal formula can be abbreviated as $[\vec{y} := \vec{\mathbf{y}}]\psi$. The conjunct of $x_1 = \mathbf{x}_1 \wedge \cdots \wedge x_k = \mathbf{x}_k$ terms is abbreviated in the same manner, i.e. $\vec{x} = \vec{\mathbf{x}}$.

The evaluation of such formulas is done using causal models with context. While the boolean connectives are interpreted as in classical propositional logic, the semantic of the additional constructs adhere to the following intuition. Firstly, a literal $x = \mathbf{x}$ holds within a particular causal model with context, if the variable x takes on the value $\mathbf{x}$ in the model. Secondly, an intervention $[\vec{Y} := \vec{\mathbf{y}}]\psi$ evaluates to true in a causal model, if the formula ψ holds in the model obtained by fixing the values of each y_i to the respective value $\mathbf{y}_i$.

Definition 4.1.10. Let $\mathcal{I}$ be a causal model, let σ be a context for $\mathcal{I}$ and let $\varphi \in \mathcal{L}_{\text{CM}}$. The relation $(\mathcal{I}, \sigma) \vDash \varphi$ is defined inductively. For

- $\varphi = (x = \mathbf{x})$, $(\mathcal{I}, \sigma) \vDash \varphi$ if the variable x has the value $\mathbf{x}$ in the unique solution to the equations in $\mathcal{I}$ given the context σ.
- $\varphi = \psi \circ \chi$ for $\circ \in \{\wedge, \vee, \rightarrow\}$ or $\varphi = \neg\psi$ then the truth value of φ is obtained as in classical propositional logic.
- $\varphi = [y_1 := \mathbf{y}_1, \ldots, y_k := \mathbf{y}_k]\psi$ then $(\mathcal{I}, \sigma) \vDash \varphi$ if $(\mathcal{I}_{y_1:=\mathbf{y}_1,\ldots,y_k:=\mathbf{y}_k}, \sigma) \vDash \psi$.

Example 4.1.5. Consider the window model $\mathcal{I} := (\Sigma, \Delta)$ from Example 4.1.2, with the context $\sigma := \{x_{AF} \mapsto \textbf{true}, x_{BF} \mapsto \textbf{false}\}$. In this model we have $(\mathcal{I}, \sigma) \vDash x_{WS} = \textbf{true}$. However, $(\mathcal{I}, \sigma) \nvDash [x_{AF} := \textbf{false}]x_{WS}$ because $(\mathcal{I}_{x_{AF}:=\textbf{false}}, \sigma) \vDash \neg x_{WS} = \textbf{false}$ due to then structural equations $x_{AF} := \textbf{false}$ and $x_{WS} := x_{AF}$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Having defined the components required for both encoding and querying causal relations, the last remaining step is to introduce Halpern's modified definition of token causality.

Definition 4.1.11 ([Hal15a])**.** $\vec{x} = \vec{\mathbf{x}}$ is an actual cause of φ in $(\mathcal{I}, \sigma)$ if

AC1: $(\mathcal{I}, \sigma) \models (\vec{x} = \vec{\mathbf{x}})$ and $(\mathcal{I}, \sigma) \models \varphi$

AC2: There are some variables $\vec{w}$ in $\mathcal{V}$ and a setting $\vec{\mathbf{x}'}$ of the variables $\vec{x}$ such that if $(\mathcal{I}, \sigma) \models (\vec{w} = \vec{\mathbf{w}})$, then $(\mathcal{I}, \sigma) \models [\vec{X} := \vec{\mathbf{x}'}, \vec{w} := \vec{\mathbf{w}}]\neg\varphi$.

AC3: $\vec{x}$ is minimal, i.e. no strict subset of $\vec{x}$ satisfies AC1 and AC2.

The intuition underlying those conditions is the following. AC1, implies that the events $(\vec{x} = \vec{\mathbf{x}})$ cannot be the cause of φ, unless both actually happen. AC3, ensures that superfluous events are not considered as causes. AC2, is by far the most demanding. The requirement of finding a setting for $\vec{x}$ such that φ does not hold, is conceptually similar to a but-for statement, i.e. but for the fact that $\vec{x} = \vec{\mathbf{x}}$, φ does not hold. Meaning, in order for a set of events to be a cause of an event, there must exists an alteration of those events under which the outcome would have changed. Furthermore, the statement fixing the values of some set of endogenous variables $\vec{w}$ amounts to a less stringent form of the ceteris paribus test. To summarise, the statements capture something akin to the following. There exist an intervention on the variables $\vec{x}$ and some set of variables $\vec{w}$ that remain fixed to their original values (regardless of the changes made to $\vec{x}$) prohibiting the event φ from occurring.

Example 4.1.6. Consider the window model $\mathcal{I} := (\Sigma, \Delta)$ from Example 4.1.2, with the context $\sigma := \{x_{AF} \mapsto \mathbf{true}, x_{BF} \mapsto \mathbf{true}\}$. We can deduce that x_{AF} is a cause of x_{WS}, i.e.

- AC1 is satisfied due to $(\mathcal{I}, \sigma) \models x_{AF}$ and $(\mathcal{I}, \sigma) \models x_{WS} = \mathbf{true}$.
- AC2 is satisfied due to $(\mathcal{I}_{x_{AF} := \mathbf{false}}, \sigma) \models \neg x_{WS} = \mathbf{true}$.
- AC3 is satisfied because no subset of $\{x_{AF}\}$ satisfies AC1 and AC2.

Notice that in AC2 the set W is empty. We could have selected $W := \{x_{BF}\}$, however, freezing this variable to the current value is immaterial as it is not affected by the intervention on x_{AF}.

Lastly, another common restriction is the restriction to boolean variables.

Definition 4.1.12 ([Hal15b])**.** Let $\mathcal{I}$ be a causal model, if $\forall x \in \mathcal{W}\, \mathcal{R}(x) = \mathbb{B}$ then $\mathcal{I}$ is a binary causal model.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

When talking about binary causal models, literals of the form $x = \mathbf{x}$ can either be $x = \mathbf{true}$ or $x = \mathbf{false}$, which can be understood as expressing that as an event either happened or not. Being essentially boolean variables, one can employ the usual notation found in logic as a shorthand. That is, the former can be abbreviated as x and the latter as $\neg x$. Similarly, an intervention on a causal model $\mathcal{I}$ can also be expressed in the same concise manner, i.e. $\mathcal{I}_{x:=\mathbf{true}}$ is simply $\mathcal{I}_x$ and $\mathcal{I}_{x:=\mathbf{false}}$ can be written as $\mathcal{I}_{\neg x}$. Moreover, a structural equation δ_x determining the value of some endogenous variable x, is in this setting a simple propositional boolean formula.

This restriction is particularly important for this thesis. Meaning that, *henceforth all causal models are considered to be binary (unless otherwise specified.)*

4.1.3 A principled approach to actual causality

Here we present `BV-CM` and `CM+T`, unless otherwise specified the definitions are taken from [BV18].

`BV-CM` uses the extension of the causal model framework `CM+T` to define its notion of token causality. `CM+T` extends binary `CM` with an additional timing function. This function maps literals into the natural numbers, which intuitively can be understood as arranging events on a timeline. Utilising those models, they define token causality by extracting a set of principles from examples. Those principles are separated into necessary and sufficient conditions. The latter essentially provide a lower bound, i.e. any definition of causality must include the necessary principles, whereas the former represent an upper bound, i.e. any definition of causality is subsumed by the sufficient principles.

For the final definition of token causality, not all principles are required. However, due to their perceived value, all principles from [BV18] will be reiterated in this subsection as well. Furthermore, rather than introducing the language extension upfront, this subsection follows the structure found in [BV18] and starts by discussing the first two of the aforementioned principles.

The first principle relies on the notion of counterfactual dependence. That is, the subsequent definition tries to express that an endogenous literal p is counterfactually dependent on an endogenous literal q if intervening on the value of q, while holding the context fixed, results in $\neg p$.

Definition 4.1.13. Given a causal model with context $(\mathcal{I}, \sigma)$ and two endogenous literals p and q such that $(\mathcal{I}, \sigma) \models p \wedge q$ then p is counterfactually dependent on q if $(\mathcal{I}_{\neg q}, \sigma) \models \neg p$.

Example 4.1.7. Consider the simple window model $\mathcal{I} := (\Sigma, \Delta)$ from Example 4.1.2, with the context $\sigma := \{x_{AF} \mapsto \mathbf{true}, x_{BF} \mapsto \mathbf{true}\}$. Because $(\mathcal{I}, \sigma) \models x_{AF} \wedge x_{WS}$ and $(\mathcal{I}_{\neg x_{AF}}, \sigma) \models \neg x_{WS}$ we know that x_{WS} is counterfactually dependent on x_{AF}. By contrast, although $(\mathcal{I}, \sigma) \models x_{BF} \wedge x_{WS}$, due to $(\mathcal{I}_{\neg x_{BF}}, \sigma) \models x_{WS}$ we know that x_{WS} is not counterfactually dependent on x_{BF}.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[BV18] argues that counterfactual dependence is a sufficient condition for causation and thus consider it as the first principle of causation.

Definition 4.1.14. Given a causal model with context $(\mathcal{I}, \sigma)$ and two endogenous literals p and q. If p is counterfactually dependent on q, then q is a cause of p w.r.t. $(\mathcal{I}, \sigma)$.

Next they introduce the notion of contributing cause, which ensures that events that fail to satisfy dependence, but are somehow responsible for the observed effect, are not neglected. For example, consider two processes that in conjunction produce an effect, yet fail to do so independently. Hence, the observed effect is not counterfactually dependent on either, however, both contributed to its existence. Anyhow, before defining the concept of contributing cause, a notion of sufficiency is required.

Definition 4.1.15. Given a causal model $\mathcal{I}$, a consistent set of literals L is sufficient for some literal p, if $p = x$ and $(\bigwedge_{l \in L} l) \rightarrow \delta_x$ or $p = \neg x$ and $(\bigwedge_{l \in L} l) \rightarrow \neg\delta_x$.

That is, a set of literals is sufficient for a positive literal, if its conjunction satisfies the structural equation determining the value of the variable contained within the literal. For a negative literal the same holds, but for the fact that the negation of the structural equation must be satisfied.

Example 4.1.8. Consider the window model $\mathcal{I} := (\Sigma, \Delta)$ from Example 4.1.2, with the context $\sigma := \{x_{AF} \mapsto \textbf{true}, x_{BF} \mapsto \textbf{true}\}$. The sets $\{x_{AF}\}$ and $\{x_{AF}, x_{BF}\}$ are sufficient for x_{WS}, because $x_{AF} \wedge x_{BF} \rightarrow x_{AF}$. By contrast, $\{x_{BF}\}$ is not.

Having defined the concept of sufficiency, one can now characterise the notion of a direct possible contributing cause, which is in fact a context independent concept.

Definition 4.1.16. Given a causal model $\mathcal{I}$ and two endogenous literals p and q. p is a direct possible contributing cause of q, if there exists a set of literals L with $p \in L$ such that L is sufficient for q, but $L \setminus \{p\}$ is not. L is called a witness for p w.r.t. q.

This notion can be generalised, to define indirect contributing causes.

Definition 4.1.17. Given a causal model $\mathcal{I}$ and two endogenous literals p and q. p is a possible contributing cause of q, if there exists a sequence of literals $p = l_1, \ldots, l_n = q$ such that $\forall i \in \{1, \ldots, n-1\}$ the literal l_i is a direct possible contributing cause of l_{i+1}.

To identify actual possible contributing causes, this definition needs to be context specific.

Definition 4.1.18. Given a causal model with context $(\mathcal{I}, \sigma)$ and two endogenous literals p and q such that $(\mathcal{I}, \sigma) \vDash p \wedge q$ then p is a direct actual contributing cause of q, if p is a direct possible contributing cause of q with a witness L such that $(\mathcal{I}, \sigma) \vDash L$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Example 4.1.9. Consider the window model $\mathcal{I} \coloneqq (\Sigma, \Delta)$ from Example 4.1.2, with context $\sigma \coloneqq \{x_{AF} \mapsto \textbf{true}, x_{BF} \mapsto \textbf{true}\}$. The literal x_{AF} is a direct possible contributing cause of x_{WS}, because $\{x_{AF}\}$ is sufficient for x_{WS}, but $\varnothing$ is not. Moreover, it is also a direct actual contributing cause, as $(\mathcal{I}, \sigma) \vDash x_{AF} \wedge x_{WS}$. By contrast, the only sufficient set for x_{WS} containing x_{BF} is $\{x_{AF}, x_{BF}\}$. However, because $\{x_{AF}\}$ remains to be sufficient, x_{BF} cannot be a direct possible contributing cause.

As before this generalises to indirect actual contributing causes.

Definition 4.1.19. Given a causal model $\mathcal{I}$ and two endogenous literals p and q such that $(\mathcal{I}, \sigma) \vDash p \wedge q$. p is an actual contributing cause of q, if there exists a sequence of literals $p = l_1, \ldots, l_n = q$ such that $\forall i \in \{1, \ldots, n-1\}$ the literal l_i is a direct actual contributing cause of l_{i+1}.

Using actual contributing causes, [BV18] finally formulate their second principle, which is a necessary condition.

Definition 4.1.20. Given a causal model $(\mathcal{I}, \sigma)$ and two endogenous literals p and q. If p is a cause of q in $(\mathcal{I}, \sigma)$, then p contributes to q w.r.t. $(\mathcal{I}, \sigma)$.

[BV18] postulate that the definition of causality must lie somewhere in between those two principles. The remaining two principles are used to demonstrate why the principle of contribution fails to be sufficient for causation. To formulate the third principle the notion of production has to be introduced. However, to do so one needs to extend the language `CM` by an additional timing function thereby creating `CM+T`.

Definition 4.1.21. A timing τ for a causal model with context $(\mathcal{I}, \sigma)$ is a function $\tau : \mathbf{L}_{(\mathcal{I},\sigma)} \to \mathbb{N}$, where $\mathbf{L}_{(\mathcal{I},\sigma)}$ is the set of all literals that hold in $(\mathcal{I}, \sigma)$, i.e. $\mathbf{L}_{(\mathcal{I},\sigma)} \coloneqq \{l \mid (\mathcal{I}, \sigma) \vDash l\}$.

Intuitively, $\tau(p) < \tau(q)$ expresses that (the event represented by) the literal p occurred before (the event represented by) the literal q, while $\tau(p) = \tau(q)$ implies that both events occurred simultaneously.

They interpret a positive literal as the occurrence of an event, while a negative literal is considered to be an omission, i.e. the absence of the event. The omissions in particular, provide a challenge w.r.t. the timing function, because we are now required to provide a timestamp for the non-occurrence of an event. They reconcile this issue by declaring that the point of time for the non-occurrence of an event is the moment in which the last event occurred that could have produced the event in question. See Definition 4.1.24 for clarification.

They generalise the notion of a timing function by allowing for partial timings, which can simplify the modelling process by allowing some events to escape the domain of τ. This is convenient because not all causal processes require keeping track of the timing. For example, the relationship between altitude and temperature is time independent.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Definition 4.1.22. A partial timing τ for a causal model with context $(\mathcal{I}, \sigma)$ is a function from some subset of the set of all literals $\mathbf{L}_{(\mathcal{I},\sigma)}$ to $\mathbb{N}$. A timing τ' extends τ, if for any literal $l \in \mathbf{L}_{(\mathcal{I},\sigma)}$, $\tau'(l) = \tau(l)$ whenever $\tau(l)$ is defined.

Adding such a new construct, interventions on causal models have to be redefined as well. Intuitively, an intervention on a timing simply assumes that everything before the intervention remains the same while everything after the intervention is unknown.

Definition 4.1.23. Given a causal model with context $(\mathcal{I}, \sigma)$, a partial timing τ and two endogenous literals p and q, such that $(\mathcal{I}, \sigma, \tau) \models p$ and $p \neq q$, the partial timing $\tau_{\neg p}$ is identical to τ up until $\tau(p) - 1$ then $\tau_{\neg p}(\neg p) = \tau(p)$ and $\tau(p) \leq \tau(q)$, τ is undefined.

The actual semantics of the binary function, as well as what restrictions such a function must adhere to, is interwoven with the stated principles upon which the token causality definition rests. One of such restrictions is the notion of being a valid timing for a particular model. Meaning that, the choice of timing is restricted since it is implicitly stated that within (strongly) recursive causal models that causes must always precede their effects. That is, if an event occurred at a certain time, the events causing it must occur before that point in time.

Definition 4.1.24. Given causal model with context and timing $(\mathcal{I}, \sigma, \tau)$, for every n let $\mathbf{L}^n_{(\mathcal{I},\sigma)} := \{l \in \mathbf{L}_{(\mathcal{I},\sigma)} \mid \tau(l) \leq n\}$. For each endogenous variable v and the literal l containing v such that $(\mathcal{I}, \sigma) \models l$, τ is considered valid for v if

- $l = v$ and $\tau(l) \geq \min_{k \in \mathbb{N}}\{\mathbf{L}^k_{(\mathcal{I},\sigma)}$ is sufficient for $l\}$;
- $l = \neg v$ and $\tau(l) = \min_{k \in \mathbb{N}}\{\mathbf{L}^k_{(\mathcal{I},\sigma)}$ is sufficient for $l\}$.

A timing is valid for $(\mathcal{I}, \sigma)$, if it is valid for all variables.

Example 4.1.10. Consider the window model $\mathcal{I} := (\Sigma, \Delta)$ from Example 4.1.2, with the context $\sigma := \{x_{AF} \mapsto \textbf{true}, x_{BF} \mapsto \textbf{true}\}$. A valid timing would be $\tau(x_{AF}) = 1$, $\tau(x_{WS}) = 2$ and $\tau(x_{BF}) = 3$. Because $\{x_{AF}\}$ is sufficient for x_{WS} and $\tau(x_{AF}) = 1$. By contrast, if given the context is $\sigma := \{x_{AF} \mapsto \textbf{false}, x_{BF} \mapsto \textbf{true}\}$ with $\tau(\neg x_{AF}) = 1$ and $\tau(x_{BF}) = 3$. Then the $\tau(x_{WS}) = 2$ would no longer be valid, i.e. the condition would require that $\tau(\neg x_{WS}) = \tau(\neg x_{AF}) = 1$.

Those restrictions naturally extend to partial timings as well.

Definition 4.1.25. A partial timing τ is possible w.r.t. $(\mathcal{I}, \sigma)$ if there exists a timing τ' that extends τ such that τ' is valid w.r.t.$(\mathcal{I}, \sigma)$.

After having sufficiently elaborated upon the notion of timing in the context of causal models, one finally move forward in defining the notion of production introduced in [BV18].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Definition 4.1.26. Given $(\mathcal{I}, \sigma, \tau)$ with τ being a valid timing for $(\mathcal{I}, \sigma)$ and two endogenous literals p and q, p is defined to be a direct producer of q if p is a direct actual contributing cause of q w.r.t. $(\mathcal{I}, \sigma)$, with a witness L such that for each $l \in L$, $\tau(l) \leq \tau(q)$.

Similar to the notion of direct contributing cause, the concept of production can be generalised to an indirect version.

Definition 4.1.27. Given $(\mathcal{I}, \sigma, \tau)$ with τ being a corresponding valid timing and let p and q be two endogenous literals. p is producer of q, if there exists a sequence of literals $p = l_1, \ldots, l_n = q$ so that for each $i \in \{1, \ldots, n-1\}$ l_i is a direct producer of l_{i+1}.

The same definition can be made for partial timings.

Definition 4.1.28. Given $(\mathcal{I}, \sigma, \tau')$ with τ' being a partial timing and two endogenous literals p and q. p is a producer of q in $(\mathcal{I}, \sigma, \tau')$ if there exists at least one valid timing τ that extends τ' such that p is a producer of q in $(\mathcal{I}, \sigma, \tau)$.

The last concept required for the third principle for causation is the notion of preemption. Meaning that while producers are literals whose contribution helped to bring about the effect, a literal that contributes, but fails to produce the effect, is classified as preempted.

Definition 4.1.29. Given $(\mathcal{I}, \sigma, \tau)$ and two endogenous literals p and q. p is preempted for q, if p contributes to q w.r.t. $(\mathcal{I}, \sigma)$ and is not a producer of p w.r.t. $(\mathcal{I}, \sigma, \tau)$.

The third principle is as follows.

Definition 4.1.30. Given $(\mathcal{I}, \sigma, \tau)$ and two endogenous literals p and q. If p is a cause of q w.r.t. $(\mathcal{I}, \sigma, \tau)$ then p is not preempted for q w.r.t. $(\mathcal{I}, \sigma, \tau)$.

This principle establishes again a necessary condition. Moreover, if taken in conjunction with the contributing principle one can strengthen the claim to p being a cause of q implying that p is a producer of q.

Corollary 4.1.0.1. *Given $(\mathcal{I}, \sigma, \tau)$ and two endogenous literals p and q. If p is a cause of q w.r.t. $(\mathcal{I}, \sigma, \tau)$ then p is a producer of q w.r.t. $(\mathcal{I}, \sigma, \tau)$.*

In stark contrast with principle number three, the fourth principle is fairly simple. Its only purpose is to regulate the relationship between the absence of a cause and the cause itself.

Definition 4.1.31. Given $(\mathcal{I}, \sigma, \tau)$ and two endogenous literals p and q. If p is a cause of q w.r.t. $(\mathcal{I}, \sigma, \tau)$ then $\neg p$ is not a cause of q w.r.t. $(\mathcal{I}_{\neg p}, \sigma, \tau_{\neg p})$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Although [BV18] define an alternative version of the fourth principle to introduce non-determinism into the definition, the previously stated ones were sufficient for their definition of their token causality. [BV18] combine the insights collected in the formalisation of the above principles to reduce the notion of token causality back to production, i.e. using said principles they were able to derive a definition of token causality using only the notion of production.

Definition 4.1.32. Given $(\mathcal{I}, \sigma, \tau)$ and two endogenous literals p and q, such that $(\mathcal{I}, \sigma, \tau) \models p \wedge q$. p is a token cause of q w.r.t. $(\mathcal{I}, \sigma, \tau)$ if p produces q and $\neg p$ does not produce q w.r.t. $(\mathcal{I}_{\neg p}, \sigma, \tau_{\neg p})$

Example 4.1.11. Consider the window model $\mathcal{I} := (\Sigma, \Delta)$ from Example 4.1.2, with the context $\sigma := \{x_{AF} \mapsto \textbf{true}, x_{BF} \mapsto \textbf{true}\}$, with the valid timing $\tau(x_{AF}) = 1$, $\tau(x_{WS}) = 2$ and $\tau(x_{BF}) = 3$. As already established in Example 4.1.9, x_{AF} is an actual contributing cause of x_{WS} and given the timing we can conclude that x_{AF} is a direct producer of x_{WS}. Now if we intervene, we obtain $(\mathcal{I}_{\neg x_{AF}}, \sigma, \tau_{\neg x_{AF}})$ with the partial timing being $\tau_{\neg x_{AF}} := \{x_{AF} \mapsto 1, x_{BF} \mapsto 3\}$. In this model we have $\neg x_{WS}$, thus $\neg x_{AF}$ can not be a producer of x_{WS} and therefore x_{AF} is a cause of x_{WS}.

4.1.4 Possible Causal Process Semantic

Here we present PCPS and CP2, unless otherwise specified the definitions are taken from [DBV19].

The token causality definition PCPS introduced in [DBV19] uses CP2, a language closely related to logic programming. Some of the defining features of CP2 are that it distinguishes between firing and enabling conditions and that a theory expressed in this language cannot contain contradicting rules. The latter condition is integral for CP2 as it allows one to elegantly make a default and deviant value distinction. Apart from that, a fairly standard condition a causal theory in PCPS must satisfy is acyclicity.

Definition 4.1.33. The language $\mathcal{L}_{\texttt{CP2}}$ over a propositional signature $\Sigma := (\mathcal{U}, \mathcal{V})$ contains only causal mechanisms $\varphi \in \mathcal{L}_{\texttt{CP2}}$ of the form

$$l \leftarrow L_T \parallel L_E$$

where

- $\leftarrow$ is the causal operator;
- $Ef(\varphi) := l$ is a literal of an endogenous variable from $\mathcal{V}$, called the effect;
- $Tc(\varphi) := L_T$ is a (possibly empty) set of literals called triggering conditions;
- $Ec(\varphi) := L_E$ is a (possibly empty) set of literals called enabling conditions.

Moreover, $Co(\varphi) := L_T \cup L_E$ is the set of conditions of φ.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The causal mechanism $l \leftarrow \|$ represents the unconditional causal mechanism causing l. Additionally, it must be noted that all literals can be partitioned into exogenous and endogenous variables, with endogenous variables being the only ones that are affected by the mechanism contained in a causal theory.

Definition 4.1.34. A causal theory Δ is a set of causal mechanisms that contains at least one mechanism for each endogenous symbol and such that:

- Δ is acyclic, i.e., there exists a strict well-founded order on symbols such that for each causal mechanism, the symbol in the effect is strictly greater than the symbols of the conditions.
- Δ does not contain mechanisms with contradictory effects, i.e. $p \leftarrow L_T \| L_E$ and $\neg p \leftarrow L'_T \| L'_E$.

The semantics accompanying this syntactic structure was constructed alongside the following intuition. Firstly, for exogenous variables there are no causal mechanisms constraining them. By contrast, for endogenous variables it is assumed that the causal theory contains all causal mechanisms affecting them. Secondly, each endogenous variable has both a default state and a deviant state, with the deviant state being the one that can be produced by activating one of the corresponding causal mechanisms. Therefore, once a variable switches into its deviant state, there is no possibility of returning to a default state. Hence, this language cannot model mechanisms such as light switches. Another consequence of this approach is that every deviant value is explained by a single active causal mechanism in the theory, while any default value, the product of inertia, can only be explained by the fact that all causal mechanisms influencing its value are blocked.

Example 4.1.12. We consider the scenario were Alice fires (AF) at a window, if Billy closed the window (BC) the window will shatter (WS). This scenario can be captured with the causal theory $\Delta := \{x_{WS} \leftarrow x_{AF} \| x_{BC}\}$, i.e. the shattering of the window is triggered by x_{AF} and enabled by x_{BC}. We will refer to this mechanism as η_1.

Due to the separation of triggering and enabling conditions, causal mechanisms have several levels of activation, namely they can be blocked, active, inactive, applicable, triggered, failed and satisfied.

Definition 4.1.35. Consider a world σ. A causal mechanism φ is

- blocked in σ by a condition $l \in Co(\varphi)$ if $\neg l \in \sigma$;
- active in world σ, if $Tc(\varphi) \subseteq \sigma$ and inactive otherwise;
- applicable in σ, if $Co(\varphi) \subseteq \sigma$;
- failed in σ, if it is active but blocked by an enabling condition;

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- satisfied in σ, if it is blocked or if its effect holds in σ.

With those distinctions in mind, one can formulate two auxiliary definitions, required for defining the structure over which statements in CP2 are evaluated.

Definition 4.1.36. Consider a world σ. Consider some endogenous literal l and some $\varphi \in \delta_l$,

- the set $\delta_l^\sigma \subseteq \delta_l$ of applicable mechanisms in σ will be called the firing set of l.
- the set $Co^{\neg\sigma}(\varphi) \subseteq Co(\varphi)$ is the set of conditions of φ that are false in σ.

Example 4.1.13. Given the scenario in Example 4.1.12 and

- the world $\sigma := \{x_{AF}, x_{BC}, x_{WS}\}$, η_1 is active due to $\{x_{AF}\} \subseteq \sigma$, applicable due to $\{x_{AF}, x_{BC}\} \subseteq \sigma$ and satisfied due to $x_{WS} \in \sigma$, thus $\delta^\sigma_{\neg x_{WS}} = \{\eta_1\}$ and $Co^{\neg\sigma}(\eta_1) = \varnothing$;
- the world $\sigma := \{\neg x_{AF}, x_{BC}, \neg x_{WS}\}$, η_1 is blocked and satisfied due to $x_{AF} \notin \sigma$, thus $\delta^\sigma_{\neg x_{WS}} = \varnothing$ and $Co^{\neg\sigma}(\eta_1) = \{\neg x_{AF}\}$;
- the world $\sigma := \{x_{AF}, \neg x_{BC}, \neg x_{WS}\}$, η_1 is active due to $\{x_{AF}\} \subseteq \sigma$ and blocked due to $x_{BC} \notin \sigma$ which implies that it is failed, thus $\delta^\sigma_{\neg x_{WS}} = \varnothing$ and $Co^{\neg\sigma}(\eta_1) = \{\neg x_{BC}\}$;
- the world $\sigma := \{\neg x_{AF}, \neg x_{BC}, \neg x_{WS}\}$, η_1 is blocked due to $x_{AF} \notin \sigma$, thus $\delta^\sigma_{\neg x_{WS}} = \varnothing$ and $Co^{\neg\sigma}(\eta_1) = \{\neg x_{AF}, \neg x_{BC}\}$.

Notice that if l is default in σ then $Co^{\neg\sigma}(\varphi) \neq \varnothing$, likewise if l is deviant in σ then $\delta_p^{\sigma_\mathcal{I}} \neq \varnothing$. A possible causal process for a causal theory Δ can be modelled as an acyclic directed graph.

Definition 4.1.37. A possible causal process for Δ over some vocabulary Σ, is a directed graph $\mathcal{I}$ with labelled vertices and edges. The set of vertices correspond to a world, denoted as $\sigma_\mathcal{I}$ and are labelled accordingly. The set of edges, all labelled with a particular mechanism, is constructed as follows. Consider some endogenous literal $p \in \sigma_\mathcal{I}$,

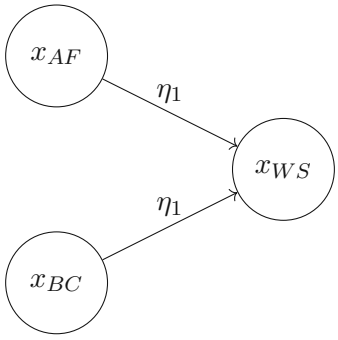
- if p is deviant, then for all $\varphi \in \delta_p^{\sigma_\mathcal{I}}$ and for all $q \in Co(\varphi)$ there exists an edge (q, p) in $E(\mathcal{I})$ which is labelled with φ. (There are no other edges to p in $E(\mathcal{I})$)
- if p is default, then for each $\varphi \in \delta_p$ and for each $q \in Co^{\neg\sigma_\mathcal{I}}(\varphi)$ here exists an edge (p, q) in $E(\mathcal{I})$ which is labelled with $\neg\varphi$. (There are no other edges to p in $E(\mathcal{I})$)

Intuitively, for a deviant endogenous literal one requires the existence of at least one path from the exogenous variable justifying the existence of the deviant value in the world. By contrast, to justify a default variable one needs to demonstrate that no causal mechanism fired.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

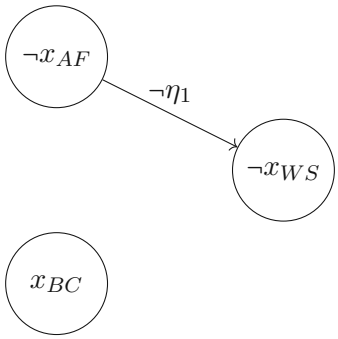
Example 4.1.14. Building on Example 4.1.13.

- Consider the world $\sigma := \{x_{AF}, x_{BC}, x_{WS}\}$ which results in



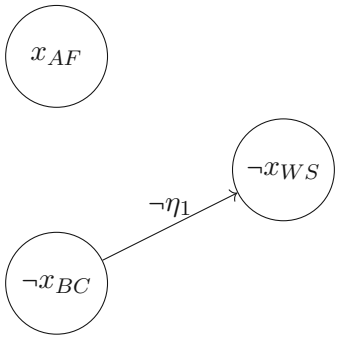
  because x_{WS} is deviant, η_1 is applicable and both x_{AF} and x_{BC} are conditions of η_1. The edge (x_{AF}, x_{WS}) is a trigger edge, while the edge (x_{BC}, x_{WS}) is an enabling edge.

- Consider the world $\sigma := \{\neg x_{AF}, x_{BC}, \neg x_{WS}\}$ which results in



  because x_{WS} is default and $Co^{\neg\sigma}(\eta_1) = \{\neg x_{AF}\}$. The edge $(\neg x_{AF}, \neg x_{WS})$ is a non-trigger edge.

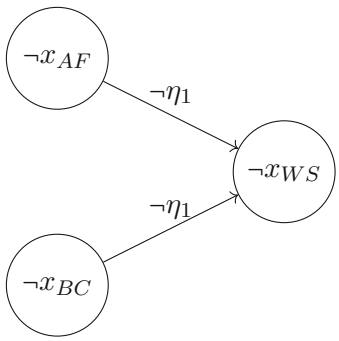
- Consider the world $\sigma := \{x_{AF}, \neg x_{BC}, \neg x_{WS}\}$ which results in



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

because x_{WS} is default and $Co^{\neg\sigma}(\eta_1) = \{\neg x_{BC}\}$. The edge $(\neg x_{BC}, \neg x_{WS})$ is a failure edge.

- Consider the world $\sigma := \{\neg x_{AF}, \neg x_{BC}, \neg x_{WS}\}$ which results in



because x_{WS} is default and $Co^{\neg\sigma}(\eta_1) = \{\neg x_{AF}, \neg x_{BC}\}$. The edge $(\neg x_{AF}, \neg x_{WS})$ is a non-trigger edge and edge $(\neg x_{BC}, \neg x_{WS})$ is a failure edge.

Definition 4.1.38. Consider some edge (q, p) in a causal process $\mathcal{I}$. If

- the edge is labelled with φ then this edge is called an active edge. Active edges are
  - trigger edges, if $q \in Tc(\varphi)$ or
  - enabling edges, if $q \in Ec(\varphi)$.
- $\neg\varphi$ then this edge is called an blocking edge. Blocking edges are
  - non-trigger edges, if $q \in Tc(\varphi)$ or
  - failure edges, if $q \in Ec(\varphi)$.

This semantic actually induces a possible world semantic.

Definition 4.1.39. A causal process $\mathcal{I}$ realises the world σ if $\sigma = \sigma_{\mathcal{I}}(= V(\mathcal{I}))$. We call σ a possible world of causal theory Δ if it is realised by some causal process for Δ.

According to [DBV19] and [DBV18] multiple notions related to token causality can be introduced using this formalism. The weakest one related to causality is the notion of influence, i.e. a variable influences another if there is a path in a causal process starting at the former and terminating at the latter.

Definition 4.1.40. A literal p is an influence of q in a possible causal process $\mathcal{I}$ of Δ, if there is a path from p to q in $\mathcal{I}$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

By simply restricting the set of causal paths used to justify the claim of influence, various refinements of the notion of influence can be defined. For example, active influence is defined, by requiring all paths to contain only active edges. The notion of token causality is stricter, allowing one to distinguish between triggering and enabling conditions, which up until now were treated symmetrically. This asymmetry ensures that enabling conditions are not considered to be token causes.

Definition 4.1.41. A literal p is an actual P-cause (production-cause) of literal q in process $\mathcal{I}$ if there is a path from p to q in $\mathcal{I}$ containing only

- trigger edges,
- non-trigger edges and
- failure edges of active causal mechanisms.

p is a direct P-cause, if there exists a path of length 1 to q, otherwise it is an indirect P-cause.

This definition implies that a path serving as a witness for P-cause claim cannot contain enabling edges and failure edges of non-active causal mechanisms. Moreover, further restrictions of this concept are possible, e.g. active P-cause.

Example 4.1.15. Building on Example 4.1.15.

- In the world $\sigma := \{x_{AF}, x_{BC}, x_{WS}\}$ we can declare x_{AF} a cause of x_{WS} because of the trigger edge. By contrast, x_{BC} cannot be a cause as it is only connected to x_{WS} by an enabling edge.
- In the world $\sigma := \{\neg x_{AF}, x_{BC}, \neg x_{WS}\}$ we can declare $\neg x_{AF}$ a cause of $\neg x_{WS}$ because of the non-trigger edge. By contrast, x_{BC} cannot be a cause as it is not connected to $\neg x_{WS}$.
- In the world $\sigma := \{x_{AF}, \neg x_{BC}, \neg x_{WS}\}$ we can declare $\neg x_{BC}$ a cause of $\neg x_{WS}$ because of the failure edge and because of η_1 being an active mechanism. By contrast, x_{AF} cannot be a cause as it is not connected to $\neg x_{WS}$.
- In the world $\sigma := \{\neg x_{AF}, \neg x_{BC}, \neg x_{WS}\}$ we can declare $\neg x_{AF}$ a cause of $\neg x_{WS}$ because of the non-trigger edge. By contrast, x_{BC} cannot be a cause, because the mechanism η_1 is not active, thus the failure edge connecting the literal to $\neg x_{WS}$ is ignored.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.1.5 Causal Inference

Here we present BCI and CT, unless otherwise specified the definitions are taken from [Boc18a].

Syntactically the language underlying Bochmans Causal Inference Theory BCI is a simple propositional language, containing an additional binary connective $\Rightarrow$, with $\varphi \Rightarrow \psi$ indicating that φ causes ψ.

Definition 4.1.42. The language $\mathcal{L}_{\mathtt{CT}}$ can be constructed over the signature $\Sigma = \mathcal{W}$ using the classic propositional language $\mathcal{L}_{\mathbb{B}}$ (over Σ), which is extended to $\mathcal{L}_{\mathtt{CT}}$ by adding

- if $\varphi \in \mathcal{L}_{\mathbb{B}}$ then $\varphi \in \mathcal{L}_{\mathtt{CT}}$.
- if $\varphi, \psi \in \mathcal{L}_{\mathbb{B}}$, then $\varphi \Rightarrow \psi \in \mathcal{L}_{\mathtt{CT}}$.

Semantically, this language can be seen as a construction consisting of two layers. The top one is non-monotonic and regulates which worlds model a certain theory, while the bottom layer specifies the behaviour of causal relation present in the language.

Definition 4.1.43. A causal theory Δ is an arbitrary set of causal rules, i.e. $\Delta \subseteq \mathcal{L}_{\mathtt{CT}}$. Moreover, let X be a set of propositions

$$\Delta(X) = \{\psi \mid \varphi \Rightarrow \psi \in \Delta \wedge \varphi \in X\}$$

is the set of propositions that are caused by X in Δ.

In the context of BCI one can speak of two different kinds of semantics, both of which are constructed using the idea of an exact model of a causal theory.

Definition 4.1.44. A consistent set of propositions X is an exact model of a causal theory Δ, if $X = Th(\Delta(X))$.

- A general nonmonotonic semantics of a causal theory Δ is the set of all its exact models.
- A causal nonmonotonic semantics of a causal theory Δ is the set of all its exact models that are worlds of Δ, i.e. that are propositional interpretations of Δ.

The semantic of this logic is non-monotonic and serves as a top layer, allowing one to plug in a suitable logic for specifying the behaviour of the causal binary connective. In [Boc18a] this relation is specified on an axiomatic basis as a production inference relation.

Definition 4.1.45. A production inference relation is a binary relation $\Rightarrow$ on the set of classical propositions satisfying the following conditions:

- *Strengthening:* If $\varphi \vDash \psi$ and $\psi \Rightarrow \chi$, then $\varphi \Rightarrow \chi$;

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- *Weakening:* If $\varphi \Rightarrow \psi$ and $\psi \vDash \chi$, then $\varphi \Rightarrow \chi$;
- *And:* If $\varphi \Rightarrow \psi$ and $\varphi \Rightarrow \chi$, then $\varphi \Rightarrow \psi \wedge \chi$;
- *Truth:* $\top \Rightarrow \top$;
- *Falsity:* $\bot \Rightarrow \bot$.

However, for the definition of token causality a production relation alone is not sufficient. Hence, the following additional properties are required.

Definition 4.1.46. Let $\Rightarrow$ be a production inference. The relation $\Rightarrow$

- is called regular, if it satisfies satisfies *Cut*, i.e. $\varphi \Rightarrow \psi$ and $\varphi \wedge \psi \Rightarrow \chi$, then $\varphi \Rightarrow \chi$.
- is called basic, if it satisfies satisfies *Or*, i.e. If $\varphi \Rightarrow \chi$ and $\psi \Rightarrow \chi$, then $\varphi \vee \psi \Rightarrow \chi$.
- is called causal, if it is basic and regular

The causal inference relation is fairly similar to the classical entailment satisfying most of its properties, but for the Reflexivity and Contraposition postulates. Moreover, since causal inference relations are basic, any rule can be rewritten to be in clausal form, i.e. $\bigwedge l_i \Rightarrow \bigvee l_j$ with l_i and l_j being classical literals. This form is essential for the definition of token causality within BCI, as token causality is (at least within BCI and some other frameworks) highly sensitive to the syntactic form of causal rules, requiring the introduction of a clausal causal theory.

Definition 4.1.47. Let Δ be a causal theory.

- Δ is called a clausal causal theory if any rule in Δ is of the form $l_1, \ldots, l_n \Rightarrow l$ (with l_i for $i \in \{1, \ldots, n\}$ and l being literals).
- Δ is called parsimonious, if no causal rule from Δ is derivable from the from the rest of the rules in Δ by causal inference.

As a quick interlude. It was demonstrated in [BL15] that the resulting language can capture binary causal models.

Definition 4.1.48. For any boolean causal model $\mathcal{I}_{\text{CM}}$, $\Delta_{\mathcal{I}_{\text{CM}}}$ is the causal theory consisting of the rules

$$\varphi \Rightarrow x \text{ and } \neg\varphi \Rightarrow \neg x$$

for δ_x in $\mathcal{I}_{\text{CM}}$ being $x := \varphi$ and

$$u \Rightarrow u \text{ and } \neg u \Rightarrow \neg u$$

for each $u \in \mathcal{U}$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Moreover, the exact worlds of $\Delta_{\mathcal{I}_{\text{CM}}}$ correspond with the solution of the structural equations from $\mathcal{I}_{\text{CM}}$, which Bochman refers to as causal worlds.

The definition of token causality presupposes some causal theory Δ and some "actual" world σ, which is in fact the exact (causal) world w.r.t. Δ.

Definition 4.1.49. Let σ be an exact world of a clausal causal theory Δ. A causal rule $l_1, \ldots, l_n \Rightarrow l$ is active in σ if $\{l_1, \ldots, l_n\} \subseteq \sigma$. Moreover, the actual sub-theory $\Delta_\sigma \subseteq \Delta$, is the set of all causal rules from Δ that are active in σ.

Additionally, Bochman introduces a relation that is dependent on the actual world.

Definition 4.1.50. Let σ be a causal world of a parsimonious clausal causal theory Δ. A literal $l' \in \sigma$ is an actual cause of a literal l in σ wrt. Δ, if and only if there exists a set of literals $L \subseteq \alpha$ such that

- $l', L \Rightarrow_\sigma l$,
- $L \not\Rightarrow_\sigma l$

with $\Rightarrow_\sigma$ being the least causal inference relation that includes Δ_σ.

To aid in understanding consider the following example.

Example 4.1.16. Consider the simple scenario described in [BV18]: If Alice fires at the window (AF), it will shatter (WS). If Bob fires at the window (BF) he will always miss. This results in the clausal causal theory $\Delta := \{x_{AF} \Rightarrow x_{WS}, \neg x_{AF} \Rightarrow \neg x_{WS}, \}$. Now given the world $\sigma := \{x_{AF}, x_{BF}\}$ the resulting sub-theory of active causal rules contains only $x_{AF} \Rightarrow x_{WS}$. Hence, x_{AF} causes x_{WS}, because $\mathbf{true} \not\Rightarrow x_{WS}$. Similarly, in the world $\sigma := \{\neg x_{AF}, x_{BF}\}$ the resulting sub-theory of active causal rules contains only $\neg x_{AF} \Rightarrow \neg x_{WS}$. Hence, $\neg x_{AF}$ causes $\neg x_{WS}$, as $\mathbf{true} \not\Rightarrow \neg x_{WS}$.

4.2 Benchmarks

In this section, we compare the previously introduced token causal definitions, i.e. `HP-15`, `BV-CM`, `PCPS` and `BCI`, by applying them to the set of benchmarks introduced in Section 3.3. These are Benchmark 3.3.1 which is an instance of *Symmetric Overdetermination*; Benchmark 3.3.2 which is an instance of *Switching*; Benchmark 3.3.3 which is an instance of *Late Preemption*; Benchmark 3.3.4 which is an instance of *Early Preemption*; Benchmark 3.3.5 which is an instance of *Double Preemption*; Benchmark 3.3.6 which is an instance of *Bogus Preemption*; Benchmark 3.3.7 which is an instance of *Short-Circuiting*.

For each of the aforementioned benchmarks, we want to establish the causes of a particular target variable. Hence, the definitions will be compared based on the set of causes they

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

identified, as well as whether those causes comply with the intuitively "correct" answer for that particular scenario.

The benchmarks Symmetric Overdetermination, Switch, Late Preemption and Early Preemption are relatively old and often discussed in the literature. Hence, we were able to find formalisations and evaluations of those scenarios for most of the discussed definitions. However, especially the newer benchmarks such as Bogus Preemption and Short Circuit are discussed less frequently and exhibit greater diversity of formalisations. Therefore, the above formalisations and results are partially taken from the literature and partially derived natively. To avoid confusion, this will be made explicit on a case by case basis.

The remaining section is structured as follows. Section 4.2.1-4.2.7 briefly re-introduce the example scenario, mention which events should be considered causes and then sketch the derivations for each of the four definitions. To accustom the reader with the inference procedures, the derivations decrease in detail as the section progresses. Lastly, Section 4.2.8 concludes by comparing the definitions given the results of the previous subsections.

4.2.1 Symmetric Overdetermination

Benchmark 3.3.1 is an instance of "Symmetric Overdetermination", which refers to the scenario, where multiple processes, all of which producing the same outcome, terminate at the same time. For a detailed discussion on this topic see Section 3.3.2. This benchmark describes the following scenario.

> Alice (AF) and Bob (BF) each fire a bullet at a window, simultaneously striking the window, shattering it (WS).

We want to establish, whether AF or BF is a cause of WS. Here the intuitive answer is that both AF and BF are causes of WS, or at the very least that both contribute to causing WS.

The modified Halpern and Pearl Definition

We define the binary causal model over the variables AF, BF and WS, i.e. $\mathcal{I} := (\Sigma, \Delta)$ for $\Sigma := (\{x_{AF}, x_{BF}\}, \{x_{WS}\}, \mathcal{R})$ with $\mathcal{R}$ being the constant function mapping to the set $\{\textbf{true}, \textbf{false}\}$, and Δ is the set of structural equations, containing the equations

$$x_{WS} := x_{AF} \vee x_{BF}$$

The story suggests the context $\sigma := \{x_{AF} \mapsto \textbf{true}, x_{BF} \mapsto \textbf{true}\}$.

Given the context we have $(\mathcal{I}, \sigma) \vDash x_{AF}$ and from the structural equation $x_{WS} := x_{AF} \vee x_{BF}$ we obtain $(\mathcal{I}, \sigma) \vDash x_{WS}$. AC1 is thus satisfied. Moreover, AC3 is trivially satisfied as the only subset is the empty set. What remains is to demonstrate AC2. However, given the context, changing the value of x_{AF} in isolation does not influence the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

value of x_{WS}, i.e. with $(\mathcal{I}, \sigma) \vDash x_{BF}$ the disjunct $x_{AF} \vee x_{BF}$ is always satisfied. Hence, x_{AF} is not a cause of x_{WS}. The same reasoning applies for x_{BF}.

Moreover, `HP-15` goes beyond mere literals as causes, thus we can check whether $x_{AF} \wedge x_{BF}$ is a cause of x_{WS}. To do so we notice that AC1 is satisfied because $(\mathcal{I}, \sigma) \vDash x_{AF} \wedge x_{BF}$ and $(\mathcal{I}, \sigma) \vDash x_{WS}$. AC2 is satisfied, as we are now allowed to modify both x_{AF} and x_{BF}, thus we can satisfy $(\mathcal{I}, \sigma) \vDash [\neg x_{AF}, \neg x_{BF}] \neg x_{WS}$. Lastly, AC3 is trivially satisfied given the observations above.

In summary, only the formula $x_{AF} \wedge x_{BF}$ is considered a cause of x_{WS}. Moreover, this result as well as the formalisation can be found in [Hal15a].

A principled approach to actual causality

We define the binary causal model with the set of structural equations containing the equations

$$x_{WS} := x_{AF} \vee x_{BF}$$

Moreover, we define the context to be $\sigma := \{x_{AF} \mapsto \textbf{true}, x_{BF} \mapsto \textbf{true}\}$ and the timing to be

$$\tau(x_{AF}) = \tau(x_{BF}) = 1 \qquad\qquad \tau(x_{WS}) = 2$$

Firstly, the timing is valid. For $\tau(x_{WS})$ we have at least one sufficient set, e.g. $\{x_{AF}\}$, which holds at time step 1. Since there is no dependence between x_{AF} and x_{BF} and since there are no other literals, assigning both the value of 1 is clearly valid.

To establish the causal claim "x_{AF} is the cause of x_{WS}", we must ensure that $(\mathcal{I}, \sigma, \tau) \vDash x_{AF} \wedge x_{WS}$, that x_{AF} produces x_{WS} w.r.t. $(\mathcal{I}, \sigma, \tau)$, and that $\neg x_{AF}$ does not produce x_{WS} w.r.t. $(\mathcal{I}_{\neg x_{AF}}, \sigma, \tau_{\neg x_{AF}})$.

Starting with the positive case, to identify a production relationship, we need to demonstrate that x_{AF} is a direct actual contributing cause of x_{WS} w.r.t. $(\mathcal{I}, \sigma)$, with a witness L such that for each $l \in L$, $\tau(l) \leq \tau(x_{WS})$. Moreover, a direct actual contributing cause requires us to find a set of literals L with $x_{AF} \in L$ such that L is sufficient for x_{WS}, but $L \setminus \{x_{AF}\}$ is not. In addition to that we require that $(\mathcal{I}, \sigma) \vDash x_{AF} \wedge x_{WS}$ holds. Lastly, L is sufficient if $(\bigwedge_{l \in L} l)$ implies $\delta_{x_{WS}}$. We propose that $\{x_{AF}\}$ is such a sufficient set, because x_{AF} clearly implies $x_{AF} \vee x_{BF}$. Moreover, x_{AF} is an actual contributing cause because $(\mathcal{I}, \sigma) \vDash x_{AF} \wedge x_{WS}$ and because $\{x_{AF}\} \setminus \{x_{AF}\}$ is not sufficient. Hence, x_{AF} is a producer of x_{WS}, because $\tau(x_{AF}) < \tau(x_{WS})$.

Now we need to consider the negative case. To do so we intervene on our model such that $\neg x_{AF}$ holds. However, $\neg x_{AF}$ does not help us to infer $x_{AF} \vee x_{BF}$, thus we can reduce the size of any sufficient set for x_{WS} by removing $\neg x_{AF}$. Hence, $\neg x_{AF}$ cannot be an actual contributing cause and therefore it cannot be a producer. Finally, allowing us to establish the claim that x_{AF} is a cause of x_{WS}.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Causal claims in this language are restricted to literals. Hence, we cannot test whether $x_{AF} \wedge x_{BF}$ is a cause of x_{WS}.

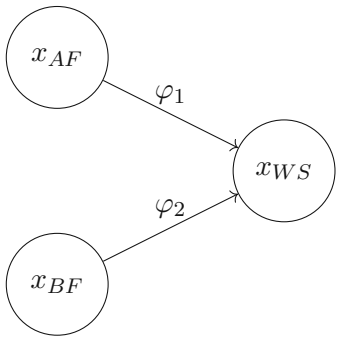
In summary, both literals x_{AF} and x_{BF} are considered causes of x_{WS}. Moreover, this result as well as the formalisation can be found in [BV18].

Possible Causal Process Semantic

We define the following causal theory

$$\varphi_1 := x_{WS} \leftarrow x_{AF} \,||$$
$$\varphi_2 := x_{WS} \leftarrow x_{BF} \,||$$

Moreover, the situation suggests the world $\sigma := \{x_{AF}, x_{BF}, x_{WS}\}$. Now notice that both causal mechanisms are applicable and active because $\{x_{AF}\} \subseteq \sigma$ and $\{x_{BF}\} \subseteq \sigma$. Furthermore, because $x_{WS} \in \sigma$ they are also satisfied. Given this we can draw the causal process $\mathcal{I}$



where both edges are trigger edges, therefore both x_{AF} and x_{BF} are actual P-causes of x_{WS}.

In summary, both literals x_{AF} and x_{BF} are considered causes of x_{WS}. Moreover, this result as well as the formalisation can be found in [DBV18].

Causal Inference

We define the following clausal causal theory

$$x_{AF} \Rightarrow x_{AF} \qquad \neg x_{AF} \Rightarrow \neg x_{AF}$$
$$x_{BF} \Rightarrow x_{BF} \qquad \neg x_{BF} \Rightarrow \neg x_{BF}$$
$$x_{AF} \Rightarrow x_{WS} \qquad \neg x_{AF}, \neg x_{BF} \Rightarrow \neg x_{WS}$$
$$x_{BF} \Rightarrow x_{WS}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

from the causal model formulation using the algorithm defined in [Boc18a]. Moreover, the situation suggests the world $\sigma := \{x_{AF}, x_{BF}, x_{WS}\}$. The resulting sub-theory of active rules is

$$x_{AF} \Rightarrow x_{AF}$$
$$x_{BF} \Rightarrow x_{BF}$$
$$x_{AF} \Rightarrow x_{WS}$$
$$x_{BF} \Rightarrow x_{WS}$$

Now given that $x_{AF} \Rightarrow x_{WS}$ and $x_{BF} \Rightarrow x_{WS}$ are active in this world and the fact that $\mathbf{true} \not\Rightarrow x_{WS}$, both x_{AF} and x_{BF} are causes of x_{WS}.

In summary, both literals x_{AF} and x_{BF} are considered causes of x_{WS}. Moreover, this result as well as the formalisation can be found in [Boc18a].

4.2.2 Switch

Benchmark 3.3.2 is an instance of Switch, which refers to the scenario, where an event serves a switch triggering one of two processes, both of which produce the same outcome. For a detailed discussion on this topic see Section 3.3.3.

This benchmark describes the following scenario.

> Alice flicks a switch (AF). The train travels on track A (TA), otherwise the train would have travelled on track B (TB). In both cases the train arrives at its destination (TD).

We want to find the causes for TD. Although there is disagreement about what the causes of TD should be. We adhere to the view that TA is a cause of TD, while both AF and TB are not.

The modified Halpern and Pearl Definition

We define the binary causal model containing the equations

$$x_{TA} := x_{AF}$$
$$x_{TB} := \neg x_{AF}$$
$$x_{TD} := x_{TA} \vee x_{TB}$$

Moreover, the story suggests the context $\sigma := \{x_{AF} \mapsto \mathbf{true}\}$.

First we want to check whether x_{AF} is a cause of x_{TD}. To answer this question we observe that $(\mathcal{I}, \sigma) \vDash x_{AF}$ and $(\mathcal{I}, \sigma) \vDash x_{TD}$. Moreover, we observe that regardless of the setting of x_{AF}, x_{TD} will always hold. However, by fixing the value of x_{TB}, which is **false** under $(\mathcal{I}, \sigma) \vDash x_{AF}$, we can obtain $(\mathcal{I}, \sigma) \vDash [\neg x_{AF} \wedge \neg x_{TB}] \neg x_{TD}$. Hence, we can declare x_{AF} a cause of x_{TD}. By contrast, $\neg x_{TB}$ is not a cause of x_{TD}, because regardless of

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

which variables we fix no intervention on x_{TB} can result in $\neg x_{TD}$. Lastly, x_{TA} is clearly a cause of x_{TD}. That is, both hold in the given model, intervening on x_{TA} results in $\neg x_{TD}$.

As a side note. This definition is able to provide the desired results by removing the intermediate variables

$$x_{TD} := x_{AF} \vee \neg x_{AF}$$

Then, $(\mathcal{I}, \sigma) \models [\neg x_{AF}]x_{TD}$ and $(\mathcal{I}, \sigma) \models [x_{AF}]x_{TD}$, which combined with the insights above indicate that x_{AF} is not a cause of x_{TD}.

In summary, both literals x_{AF} and x_{TA} are considered causes of x_{TD}. In this particular instance we used the formalisation found in [Wes15]. Hence, the derived results are independent of the discussion found in [Hal15a].

A principled approach to actual causality

We define the binary causal model containing the equations

$$x_{TA} := x_{AF}$$
$$x_{TB} := \neg x_{AF}$$
$$x_{TD} := x_{TA} \vee x_{TB}$$

Moreover, we define the context to be $\sigma := \{u_{AF} \mapsto \textbf{true}\}$ and since timing is not important we simply set τ to be the constant function 1, as we only require that causes happen either before or at the same moment as their effects.

We test whether x_{AF} is a cause of x_{TD}. To establish production we notice that $L := \{x_{AF}\}$ is a sufficient set, whereas the empty set is not. Hence, x_{AF} is an actual contributing cause of x_{TA}. Moreover, since τ is constant and $|L| = 1$ it suffices to notice that $\tau(x_{AF}) \leq \tau(x_{TA})$, in order to establish that x_{AF} directly produces x_{TA}. Moreover, by the same reasoning we obtain that x_{TA} directly produces x_{TD}. Hence, we obtain that x_{AF} produces x_{TD}.

However, notice that $\neg x_{AF}$ is a producer of x_{TD} in the modified model $(\Delta_{\neg x_{AF}}, \sigma, \tau_{\neg x_{AF}})$. That is, we have $(\Delta_{\neg x_{AF}}, \sigma, \tau_{\neg x_{AF}}) \models x_{TB}$. Now $\{\neg x_{AF}\}$ trivially implies $\neg x_{AF}$ while the empty set does not. Hence, with a constant timing we obtain that $\neg x_{AF}$ directly produces x_{TB}. By similar reasoning we obtain x_{TB} directly produces x_{TD}, which allows us to deduce that $\neg x_{AF}$ produces x_{TD}. Hence, we can conclude that neither x_{TA} nor $\neg x_{TA}$ are causes of x_{TD}.

Having already established that x_{TA} produces x_{TD}, we only need to show that $\neg x_{TA}$ does not produce x_{TD} w.r.t. $(\Delta_{\neg x_{TA}}, \sigma, \tau_{\neg x_{TA}})$. However, it is easy to see that x_{TD} does not hold in $(\Delta_{\neg x_{TA}}, \sigma, \tau_{\neg x_{TA}})$, which implies that it cannot be produced by $\neg x_{TA}$ and we obtain x_{TA} is a cause of x_{TD}.

Finally, to show that $\neg x_{TB}$ is not a cause of x_{TD}, it is sufficient to notice that in the original model, any sufficient set containing $\neg x_{TB}$ remains sufficient once this literal is removed. Hence, $\neg x_{TB}$ does not produce x_{TD} and thus fails to be a cause.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

In summary, only the literal x_{TA} is considered a cause of x_{TD}. The formalisation and the result that x_{AF} is not a cause of x_{TD}, can be found in [BV18]. However, the remaining literals had to be checked independently.

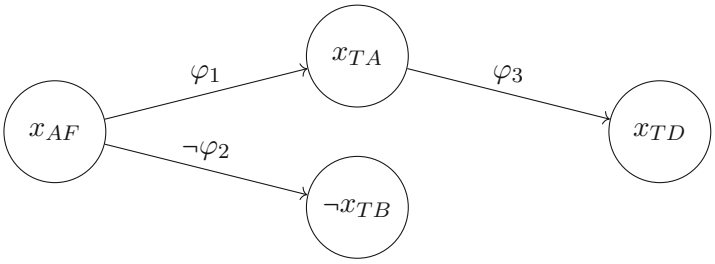
Possible Causal Process Semantic

We define the following causal theory

$$\begin{aligned}\varphi_1 &:= x_{TA} \leftarrow \| \, x_{AF} \\ \varphi_2 &:= x_{TB} \leftarrow \| \, \neg x_{AF} \\ \varphi_3 &:= x_{TD} \leftarrow x_{TA} \, \| \\ \varphi_4 &:= x_{TD} \leftarrow x_{TB} \, \|\end{aligned}$$

Moreover, the situation suggests the world $\sigma := \{x_{AF}, x_{TA}, \neg x_{TB}, x_{TD}\}$. Now notice that φ_1 and φ_3 are applicable, active and satisfied. φ_2 is a failed causal mechanism, i.e. it is active and blocked by the enabling condition $\neg x_{AF}$. Finally, φ_4 is blocked and satisfied.

Given this we can draw the causal process $\mathcal{I}$



where the edge (x_{TA}, x_{TD}) is a trigger edge, therefore x_{TA} is an actual P-cause of x_{TD}. However, since x_{AF} is only an enabling condition in φ_1 the edge (x_{AF}, x_{TA}) is an enabling edge and thus x_{AF} is not an actual P-cause of x_{TD}. Moreover, because φ_4 is not applicable, $\neg x_{TB}$ cannot be an actual P-cause of x_{TD}.

In summary, only the literal x_{TA} is considered a cause of x_{TD}. The formalisation and the result that x_{AF} is not a cause of x_{TD}, can be found in [DBV18] in a slightly different form. However, the remaining literals had to be checked independently.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Causal Inference

We define the following clausal causal theory

$$\begin{aligned} x_{AF} &\Rightarrow x_{TA} & \neg x_{AF} &\Rightarrow \neg x_{TA} \\ \neg x_{AF} &\Rightarrow x_{TB} & x_{AF} &\Rightarrow \neg x_{TB} \\ x_{TA} &\Rightarrow x_{TD} & \neg x_{TA}, \neg x_{TA} &\Rightarrow \neg x_{TD} \\ x_{TB} &\Rightarrow x_{TD} & & \\ x_{AF} &\Rightarrow x_{AF} & \neg x_{AF} &\Rightarrow \neg x_{AF} \end{aligned}$$

Moreover, the situation suggests the world $\sigma := \{x_{AF}, x_{TA}, \neg x_{TB}, x_{TD}\}$, thus the following causal rules are active.

$$\begin{aligned} x_{AF} &\Rightarrow x_{AF} \\ x_{AF} &\Rightarrow x_{TA} \\ x_{TA} &\Rightarrow x_{TD} \\ x_{AF} &\Rightarrow \neg x_{TB} \end{aligned}$$

From $x_{AF} \Rightarrow x_{TA}$ and $x_{TA} \Rightarrow x_{TD}$ we get $x_{AF} \Rightarrow x_{TD}$ and because of $\mathbf{true} \not\Rightarrow x_{TD}$ we obtain x_{AF} is a cause of x_{TD}. Moreover, due to $x_{TA} \Rightarrow x_{TD}$ we have x_{TA} is a cause of x_{TD} given that $\mathbf{true} \Rightarrow x_{TD}$. Finally, given the sub-theory we cannot derive that x_{TB} is a cause of x_{TD}.

In summary, both literals x_{TA} and x_{AF} are considered causes of x_{TD}. The formalisation and the results can be found in [Boc18a] in a slightly different form.

4.2.3 Late Preemption

Here we test Benchmark 3.3.3 against the presented formalisms. This example is an instance of "Late Preemption", which refers to the scenario, where there are two causal processes running in parallel, both would produce the same outcome, but one process terminates before the other does. For a detailed discussion on this topic see Section 3.3.4.

This benchmark describes the following scenario.

> Alice (AF) and Bob (BF) each fire a bullet at a window. Alice's bullet hits the window first (AH). The window shatters (WS). Bob's bullet arrives second and does not hit the window (BH).

We want to find the causes for WS. Here intuition dictates that AF and AH are causes of WS, while BF and BH are not.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The modified Halpern and Pearl Definition

We define the binary causal model containing the equations

$$x_{AH} := x_{AF}$$
$$x_{BH} := x_{BF} \wedge \neg x_{AH}$$
$$x_{WS} := x_{AH} \vee x_{BH}$$

Moreover, the story induced the context $\sigma := \{x_{AF} \mapsto \textbf{true}, x_{BF} \mapsto \textbf{true}\}$.

First, x_{AF} is a cause of x_{WS}. That is, if we freeze the value of x_{BH}, which under the current model evaluates to false, then we obtain $\neg x_{WS}$, i.e. $(\mathcal{I}, \sigma) \models [\neg x_{AF} \wedge \neg x_{BH}]\neg x_{WS}$. Due to x_{AF} being essentially a proxy for x_{AH}, the same argument can be employed to establish x_{AH} is a cause of x_{WS}.

x_{BF} fails to be a cause of x_{WS}, because $x_{BF} \wedge \neg x_{AH}$ will always be false regardless of the value of x_{BF} as we are unable to modify x_{AH}. Moreover, $\neg x_{BH}$ fails to be a cause as well, due to $x_{AH} \vee x_{BH}$ being true regardless of the value of x_{BH}.

In summary, both literals x_{AF} and x_{AH} are considered causes of x_{WS}. Moreover, the formalisation and the results can be found in [Hal15a].

A principled approach to actual causality

We define the binary causal model containing the equations

$$x_{AH} := x_{AF}$$
$$x_{BH} := x_{BF}$$
$$x_{WS} := x_{AH} \vee x_{BH}$$

Notice, that because of the timing introduced below, we can replace the structural equation for x_{BH} with $x_{BH} := x_{BF}$. Moreover, we define the context to be $\sigma := \{x_{AF} \mapsto \textbf{true}, x_{BF} \mapsto \textbf{true}\}$ and the timing to be

$$\tau(x_{AF}) := \tau(x_{BF}) = 1$$
$$\tau(x_{AH}) := 2$$
$$\tau(x_{WS}) := 3$$
$$\tau(x_{BH}) := 4$$

Moreover, because all literals satisfied are positive and all causes come before their effects, the timing is valid.

It is easy to see that x_{AF} is a direct producer of x_{AH} because of $\tau(x_{AF}) \leq \tau(x_{AH})$. In a similar fashion, x_{AH} is a direct producer of x_{WS} because of $\tau(x_{AH}) \leq \tau(x_{WS})$ and because $x_{AH} \rightarrow x_{AH} \vee x_{BH}$ holds, while $\top \rightarrow x_{AH} \vee x_{BH}$ does not. Hence, x_{AF} produces x_{WS}. By contrast, $\neg x_{AF}$ does not produce x_{WS} in $(\Delta_{\neg x_{AF}}, \sigma, \tau_{\neg x_{AF}})$, because $\neg x_{AF}$ produces neither x_{AH} nor x_{BH}. Therefore, failing to produce x_{WS}, implying that x_{AF} is

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

in fact a cause. Moreover, since $\neg x_{AH}$ does not produce x_{WS} in $(\Delta_{\neg x_{AH}}, \sigma, \tau_{\neg x_{AH}})$ we obtain that x_{AH} is a cause as well.

As for x_{BF}, while it produces x_{BH} in a symmetric fashion to x_{AF} and x_{AH}. The chain of production fails because x_{BH} is a mere actual contributing cause of x_{WS}. That is, while $(\Delta, \sigma, \tau) \vDash x_{BH} \wedge x_{WS}$ and $\{x_{BH}\}$ is sufficient for x_{WS} while $\varnothing$ is not, we have $\tau(x_{BH}) = 4 > 3 = \tau(x_{WS})$. Hence, any sufficient set that contains x_{BH} fails to be a proper witness to establish production.

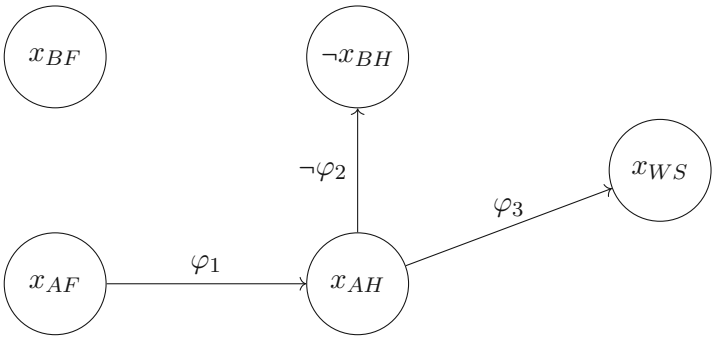
In summary, both literals x_{AF} and x_{AH} are considered causes of x_{WS}. Moreover, the formalisation and the results can be found in [BV18].

Possible Causal Process Semantic

We define the following causal theory

$$\begin{aligned}
\varphi_1 &:= x_{AH} \leftarrow x_{AF} \parallel \\
\varphi_2 &:= x_{BH} \leftarrow x_{BF} \parallel \neg x_{AH} \\
\varphi_3 &:= x_{WS} \leftarrow x_{AH} \parallel \\
\varphi_4 &:= x_{WS} \leftarrow x_{BH} \parallel
\end{aligned}$$

Moreover, the situation suggests the world $\sigma := \{x_{AF}, x_{BF}, x_{AH}, \neg x_{BH}, x_{WS}\}$. Now notice that φ_1 and φ_3 are applicable, active and satisfied, while φ_2 is a failed causal mechanism, i.e. it is active and blocked by the enabling condition $\neg x_{AH}$. Given $\neg x_{BH}$ the mechanism φ_4 is blocked but satisfied. Given this we can draw the causal process $\mathcal{I}$.



The edges (x_{AF}, x_{AH}) and (x_{AH}, x_{WS}) are trigger edges, thus both x_{AF} and x_{AH} are causes of x_{WS}. By contrast, there does not exist a path that reaches x_{WS} from x_{BF} and $\neg x_{BH}$, thus both fail to be causes.

In summary, both literals x_{AF} and x_{AH} are considered causes of x_{WS}. Moreover, the formalisation and the results can be found in [BV18].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Causal Inference

We define the following clausal causal theory

$$
\begin{aligned}
x_{AF} &\Rightarrow x_{AF} & \neg x_{AF} &\Rightarrow \neg x_{AF} \\
x_{BF} &\Rightarrow x_{BF} & \neg x_{BF} &\Rightarrow \neg x_{BF} \\
x_{AF} &\Rightarrow x_{AH} & \neg x_{AF} &\Rightarrow \neg x_{AH} \\
x_{BF}, \neg x_{AF} &\Rightarrow x_{BH} & \neg x_{BF} &\Rightarrow \neg x_{BH} \\
x_{AH} &\Rightarrow x_{WS} & x_{AF} &\Rightarrow \neg x_{BH} \\
x_{BH} &\Rightarrow x_{WS} & \neg x_{AB}, \neg x_{BH} &\Rightarrow \neg x_{WS}
\end{aligned}
$$

Moreover, the situation suggests the world $\sigma := \{x_{AF}, x_{BF}, x_{AH}, \neg x_{BH}, x_{WS}\}$, thus the following causal rules are active.

$$
\begin{aligned}
x_{AF} &\Rightarrow x_{AF} \\
x_{BF} &\Rightarrow x_{BF} \\
x_{AF} &\Rightarrow x_{AH} \\
x_{AH} &\Rightarrow x_{WS} \\
x_{AF} &\Rightarrow \neg x_{BH}
\end{aligned}
$$

From $x_{AF} \Rightarrow x_{AH}$ and $x_{AH} \Rightarrow x_{WS}$ we get $x_{AF} \Rightarrow x_{WS}$ and because of $\mathbf{true} \not\Rightarrow x_{WS}$ we obtain x_{AF} and x_{AH} are causes of x_{WS}. Moreover, given this sub-theory we are unable to infer that $\neg x_{BH}$ and x_{BF} are causes of x_{WS}.

In summary, both literals x_{AF} and x_{AH} are considered causes of x_{WS}. Moreover, the formalisation and the results can be found in [Boc18a].

4.2.4 Early Preemption

Benchmark 3.3.4 is an instance of "Early Preemption", which refers to the scenario, where there are two causal processes, both would produce the same outcome, but one process terminates before the other can even start. For a detailed discussion on this topic see Section 3.3.5.

Benchmark 3.3.4 describes the following scenario.

> Alice fires a bullet at the window (AF). If Alice hits the window, the window shatters (WS). If Alice does not hit the window, Bob fires a bullet at the window (BF), hitting it (BH) leading to its shattering.

We want to find the causes for WS. Here it is slightly unclear whether AF should be considered a cause of WS, see the discussion in Section 3.3.5. However, in this instance we side with the majority and declare that AF should be considered a cause while $\neg BF$ should not.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The modified Halpern and Pearl Definition

We define the binary causal model containing the equations

$$x_{BF} := \neg x_{AF}$$
$$x_{WS} := x_{AF} \vee x_{BF}$$

Moreover, the story induced the context $\sigma := \{x_{AF} \mapsto \textbf{true}\}$.

First, x_{AF}. If we freeze the value of x_{BF}, which under the current model amounts to false, i.e. $(\mathcal{I}, \sigma) \models \neg x_{BF}$. Then we obtain $(\mathcal{I}, \sigma) \models [\neg x_{AF} \wedge \neg x_{BH}]\neg x_{WS}$. Hence, x_{AF} is a cause of x_{WS}. Second, $\neg x_{BF}$. It fails to be a cause, because the value of x_{WS} will always be true regardless of the value of x_{BF} as we are unable to modify x_{AF}.

In summary, only the literal x_{AF} is considered a cause of x_{WS}. Moreover, the discussed formalisation of the scenario was taken from [BV18]. Hence, the discussed results had to be derived independently of the literature.

A principled approach to actual causality

We define the binary causal model containing the equations

$$x_{BF} := \neg x_{AF}$$
$$x_{WS} := x_{AF} \vee x_{BF}$$

Moreover, we define the context to be $\sigma := \{u_{AF} \mapsto \textbf{true}\}$ and the timing to be

$$\tau(x_{AF}) := 1$$
$$\tau(\neg x_{BF}) := 1$$
$$\tau(x_{WS}) := 2$$

Notice that it would have been sufficient to set τ to be constant. However, it would not have been valid to set $\tau(\neg x_{BF}) > 1$, because x_{AF} occurs at time 1 and after that no event can bring forth x_{BF}.

We want to assess whether x_{AF} is a cause of x_{WS}. We can observe that this behaves similar to a Switch scenario. To be precise, x_{AF} is an actual contributing cause, because $\{x_{AF}\}$ is sufficient and the empty set is not. Now with $\tau(x_{AF}) \leq \tau(x_{WS})$ we establish production. However, $\neg x_{AF}$ produces x_{BF} in $(\Delta_{\neg x_{AF}}, \sigma, \tau_{\neg x_{AF}})$, because $\{\neg x_{AF}\}$ is sufficient for x_{BF} while $\varnothing$ is not and there exists at least one extension of the partial timing $\tau_{\neg x_{AF}} := \{x_{AF} \mapsto 1\}$ such that $\tau_{\neg x_{AF}}(\neg x_{AF}) \leq \tau_{\neg x_{AF}}(x_{BS})$. An argument similar to the one for x_{AF} can be made to establish that x_{BF} produces x_{WS}. Hence, we obtain $\neg x_{AF}$ produces x_{WS} and must therefore conclude that x_{AF} is not a cause of x_{WS}. Moreover, $\neg x_{BF}$ fails to be a cause of x_{WS}, because there cannot be a minimally sufficient set containing $\neg x_{BF}$.

In summary, we can conclude there is no cause of x_{WS}. In [BV18] they discuss a more involved formalisation that includes the accuracy of both participants. Hence, the exact

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

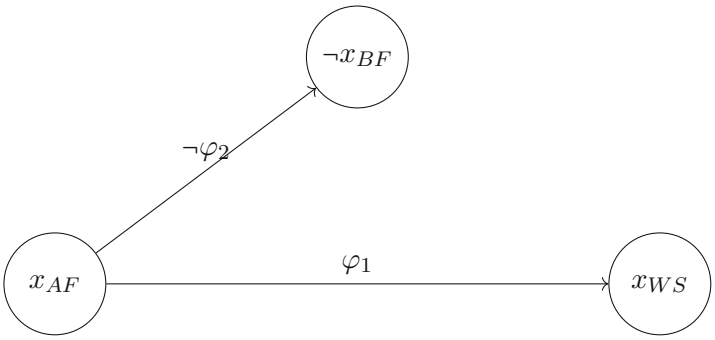
results are not present in [BV18]. However, given that they consider Early Preemption as an instance of Switch, the derived results are similar to the Switch-example found in [BV18].

Possible Causal Process Semantic

We define the following causal theory

$$\varphi_1 := x_{WS} \leftarrow x_{AF} \,||$$
$$\varphi_2 := x_{BF} \leftarrow \neg x_{AF} \,||$$
$$\varphi_3 := x_{WS} \leftarrow x_{BF} \,||$$

Moreover, the situation suggests the world $\sigma := \{x_{AF}, \neg x_{BF}, x_{WS}\}$. Now notice that φ_1 is applicable, active and satisfied. φ_2 and φ_3 are blocked and satisfied. Given this we can draw the causal process $\mathcal{I}$



where the edge labelled with φ_1 is a trigger edge, therefore x_{AF} is an actual P-cause of x_{WS}. Because, φ_3 is not applicable, $\neg x_{BF}$ is not an actual P-cause of x_{WS}.

In summary, only the literal x_{AF} is considered a cause of x_{WS}. The formalisation and the results can be found in [DBV18] in a slightly different form.

Causal Inference

Similar as it is done in [Boc18a], we define the following clausal causal theory

$$x_{AF} \Rightarrow x_{AF} \qquad \neg x_{AF} \Rightarrow \neg x_{WS}$$
$$\neg x_{AF} \Rightarrow x_{BF} \qquad x_{AF} \Rightarrow \neg x_{BF}$$
$$x_{AF} \Rightarrow x_{WS} \qquad \neg x_{AF}, \neg x_{BF} \Rightarrow \neg x_{WS}$$
$$x_{BF} \Rightarrow x_{WS}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Moreover, the situation suggests the world $\sigma := \{x_{AF}, \neg x_{BF}, x_{WS}\}$, thus the following causal rules are active.

$$x_{AF} \Rightarrow x_{AF}$$
$$x_{AF} \Rightarrow x_{WS}$$
$$x_{AF} \Rightarrow \neg x_{BF}$$

From $x_{AF} \Rightarrow x_{WS}$ and $\mathbf{true} \not\Rightarrow x_{WS}$ we obtain x_{AF} is a cause of x_{WS}. Moreover, we cannot derive that x_{BF} causes x_{WS}.

In summary, the literal x_{AF} is a cause of x_{WS}. Moreover, this result as well as the formalisation can be found in [Boc18a].

4.2.5 Double Preemption

Benchmark 3.3.5 is an instance of "Double Preemption", which refers to the scenario, where a process that would have prevented another process, was prevented by an entirely different process itself. For a detailed discussion on this topic see Section 3.3.6.

This benchmark describes the following scenario.

> Alice intends to fire a bullet at a window (AI). Bob intends to prevent Alice from hitting the window (BI). Carol intends to prevent Bob from stopping Alice (CI). Bob tries to stop Alice (BSA). Bob is stopped by Carol (CSB). Alice fires a bullet (AF), hits the window (AH) and shatters it (WS). The window shatters (WS).

We want to identify the cause of WS. According to [Hal16a, p. 35], AI, AF, CI, CSB and $\neg BSA$ should be considered causes.

The modified Halpern and Pearl Definition

We define the binary causal model Δ containing the equations

$$x_{CSA} := x_{CI}$$
$$x_{BSA} := x_{BI} \wedge \neg x_{CSB}$$
$$x_{AF} := x_{AI} \wedge \neg x_{BSA}$$
$$x_{WS} := x_{AF}$$

Moreover, the story induced the context $\sigma := \{x_{AI} \mapsto \mathbf{true}, x_{BI} \mapsto \mathbf{true}, x_{CI} \mapsto \mathbf{true}\}$.

First, x_{AI} is a cause of x_{WS}, because we simply need to intervene such that x_{AI} is false and without freezing any other value we obtain $(\mathcal{I}, \sigma) \models [\neg x_{AI}]\neg x_{WS}$. The same holds for x_{AF}. Second, x_{BI} is not a cause of x_{WS}, because we cannot intervene such that x_{CSB} does not hold, thus regardless of the value of x_{BI} the value of x_{BSA} will remain the same.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Therefore, no form of intervention on x_{BI} can influence x_{CSB}. By contrast, $\neg x_{BSA}$ is a cause of x_{WS}, because if we intervene on x_{BSA} by setting it to true, we observe that x_{AF} and subsequently x_{WS} will evaluate to false. Third, x_{CI} is a cause of x_{WS}. To see this we simply intervene such that $\neg x_{CI}$ holds. After this intervention x_{BSA} is now satisfied, which leads to $\neg x_{AF}$ and thus directly to $\neg x_{WS}$, i.e. $(\mathcal{I}, \sigma) \models [\neg x_{CI}]\neg x_{WS}$. A similar argument allows us to claim that x_{CSB} is a cause as well.

In summary, the literals x_{AI}, x_{AF}, x_{CI}, x_{CSB} and $\neg x_{BSA}$ and x_{AH} are considered causes of x_{WS}. This formalisation differs from the one found in [Hal16a] in some immaterial aspects. Hence, most of the results are taken from the literature.

A principled approach to actual causality

We define the binary causal model Δ containing the equations

$$\begin{aligned} x_{CSA} &:= x_{CI} \\ x_{BSA} &:= x_{BI} \wedge \neg x_{CSB} \\ x_{AF} &:= x_{AI} \wedge \neg x_{BSA} \\ x_{WS} &:= x_{AF} \end{aligned}$$

Moreover, we define the context to be $\sigma := \{x_{AI} \mapsto \textbf{true}, x_{BI} \mapsto \textbf{true}, x_{CI} \mapsto \textbf{true}\}$ and to keep things simple we define the timing to be the constant function 1.

First, x_{AF} is a cause of x_{WS}, i.e. any sufficient set for x_{WS} must contain x_{AF}, now given that both hold in our model and that the timing is constant it follows that x_{AF} is a producer of x_{WS}. By contrast, $\neg x_{AF}$ cannot be such a producer, as x_{WS} does not hold in the modified model.

Second, x_{AI} is a cause of x_{WS}. To establish this we observe that the set $\{x_{AI}, \neg x_{BSA}\}$ is sufficient for x_{AF}, while $\{\neg x_{BSA}\}$ is not. Moreover, we know that $\neg x_{BSA} \wedge x_{AI}$ holds, as x_{CI} implies x_{CSA} which blocks us from inferring x_{BSA}. Hence, we have established x_{AF} to be an actual contributing cause, which given a constant timing establishes production. Moreover, it is easy to see that x_{AF} produces x_{WS}. This allows us to claim causal behaviour, as $\neg x_{AI}$ cannot be an actual contributing cause of x_{AF} in $(\Delta_{\neg x_{AI}}, \sigma, \tau_{\neg x_{AI}})$ because x_{AF} cannot be inferred in this model.

Third, $\neg x_{BSA}$ is a cause of x_{WS}. This follows a similar argument as x_{AI}. That is, $\{x_{AI}, \neg x_{BSA}\}$ is a sufficient set, while $\{x_{AI}\}$ is not. Now given that both $\neg x_{BSA}$ and x_{AF} hold in the model and that the timing is constant we obtain that $\neg x_{BSA}$ produces x_{AF} which then produces x_{WS}. By contrast, x_{BSA} cannot produce x_{AF}.

Fourth, x_{BI} is a not cause of x_{WS}. This follows from the observation that the sequence of direct producers for x_{WS} requires x_{BSA} to hold, which as already established cannot be derived. Hence, the condition $(\Delta, \sigma, \tau) \models x_{BI} \wedge x_{BSA}$ is not satisfied.

Fifth, x_{CSA} is a cause of x_{WS}, because the set $\{x_{CSA}\}$ is sufficient for $\neg x_{BSA}$, i.e. x_{CSA} implies $\neg(x_{BI} \wedge \neg x_{CSB})$. Moreover, given a constant timing and the fact that the empty

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

set is insufficient, we can claim that x_{CSA} produces $\neg x_{BSA}$. Following the chain of production results in x_{CSA} produces x_{WS}. As for the negative case. If we intervene on x_{CSA} by setting it to false, the resulting model will satisfy x_{BSA} and thus x_{WS} cannot hold.

Lastly, x_{CI} is a cause of x_{WS}. The positive case is easy to see, while for the negative it suffices to notice that x_{CSA} does not hold in the modified model.

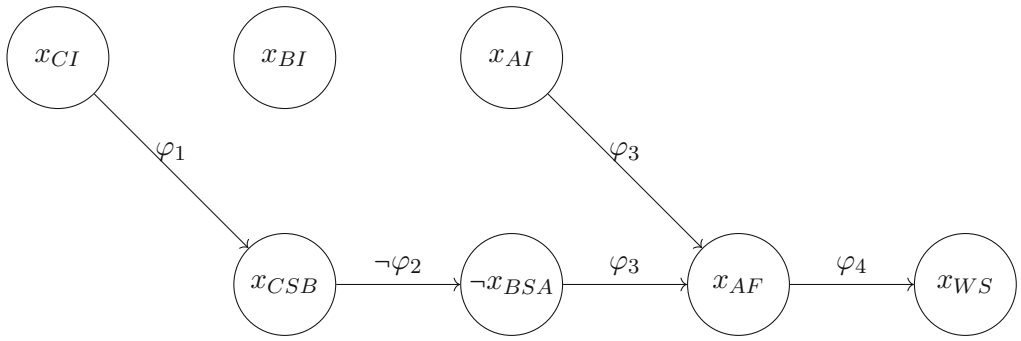
In summary, the literals x_{AI}, x_{AF}, x_{CI}, x_{CSB} and $\neg x_{BSA}$ and x_{AH} are considered causes of x_{WS}. A more concise version of this benchmark, as well as the respective results, can be found in [BV18].

Possible Causal Process Semantic

We define the following causal theory

$$\begin{aligned}
\varphi_1 &:= x_{CSA} \leftarrow x_{CI} \parallel \\
\varphi_2 &:= x_{BSA} \leftarrow x_{BI} \parallel \neg x_{CSB} \\
\varphi_3 &:= x_{AF} \leftarrow x_{AI} \parallel \neg x_{BSA} \\
\varphi_4 &:= x_{WS} \leftarrow x_{AF} \parallel
\end{aligned}$$

Moreover, the situation suggests the world $\sigma := \{x_{AI}, x_{BI}, x_{CI}, x_{CSA}, \neg x_{BSA}, x_{AF}, x_{WS}\}$. Now notice that φ_1, φ_3 and φ_4 are applicable, active and satisfied. By contrast, φ_3 is a failed causal mechanism. Given this we can draw the causal process $\mathcal{I}$



Notice that the edge (x_{BSA}, x_{AF}) is an enabling edge. Hence, only x_{AI} and x_{AF} are actual P-causes of x_{WS}.

In summary, only the literals x_{AI} and x_{AF} are considered causes of x_{WS}. A more concise version formalisation of this benchmark, as well as the respective results can be found in [DBV18].

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Causal Inference

We define the following clausal causal theory

$$
\begin{aligned}
x_{AI} &\Rightarrow x_{AI} & \neg x_{AI} &\Rightarrow \neg x_{AI}\\
x_{BI} &\Rightarrow x_{BI} & \neg x_{BI} &\Rightarrow \neg x_{BI}\\
x_{CI} &\Rightarrow x_{CI} & \neg x_{CI} &\Rightarrow \neg x_{CI}\\
x_{CI} &\Rightarrow x_{CSB} & \neg x_{CI} &\Rightarrow \neg x_{CSA}\\
x_{BI}, \neg x_{CSA} &\Rightarrow x_{BSA} & \neg x_{BI} &\Rightarrow \neg x_{BSA}\\
& & x_{CSB} &\Rightarrow \neg x_{BSA}\\
x_{AI}, \neg x_{BSA} &\Rightarrow x_{AF} & \neg x_{AI} &\Rightarrow \neg x_{AF}\\
& & x_{BSA} &\Rightarrow \neg x_{AF}\\
x_{AF} &\Rightarrow x_{WS} & \neg x_{AF} &\Rightarrow \neg x_{WS}
\end{aligned}
$$

Moreover, the situation suggests the world $\sigma := \{x_{AI}, x_{BI}, x_{CI}, x_{CSB}, \neg x_{BSA}, x_{AF}, x_{WS}\}$, thus the following causal rules are active.

$$
\begin{aligned}
x_{AI} &\Rightarrow x_{AI}\\
x_{BI} &\Rightarrow x_{BI}\\
x_{CI} &\Rightarrow x_{CI}\\
x_{CI} &\Rightarrow x_{CSB}\\
x_{AI}, \neg x_{BSA} &\Rightarrow x_{AF}\\
x_{AF} &\Rightarrow x_{WS}\\
x_{CSB} &\Rightarrow \neg x_{BSA}
\end{aligned}
$$

First, x_{AF} is a cause of x_{WS} because of $x_{AF} \Rightarrow x_{WS}$ and $\mathbf{true} \not\Rightarrow x_{WS}$. Moreover, by transitivity we obtain $x_{AI}, \neg x_{BSA} \Rightarrow x_{WS}$ and given that $\neg x_{BSA} \neg \Rightarrow x_{AF}$ it follows that x_{AI} is a cause as well. Similarly, we have $\neg x_{BSA}$ is a cause of x_{WS}. Second, x_{CSB} causes x_{WS} by transitivity, because of $x_{CSB} \Rightarrow \neg x_{BSA}$ but $\mathbf{true} \not\Rightarrow \neg x_{BSA}$. From this it is easy to see that x_{CI} causes x_{WS}. By contrast, x_{BI} is not contained in our sub-theory. Hence, we can not derive x_{BI} causes x_{WS}.

In summary, the literals x_{AI}, x_{AF}, x_{CI}, x_{CSB} and $\neg x_{BSA}$ and x_{AH} are considered causes of x_{WS}. The formalisation is obtained by translating the causal model introduced above using the algorithm in [Boc18a]. Moreover, we have not found a suitable formalisation of this benchmark in the literature.

4.2.6 Bogus Preemption

Benchmark 3.3.6 is an instance of "Bogus Preemption", which refers to the scenario, where when an action is taken to interrupt an inactive process. For a detailed discussion on this topic see Section 3.3.7.

We consider the extended version of Bogus Preemption described in Benchmark 3.3.6, i.e.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

> Alice intents to put lethal poison into Carol's water. However, Alice does not put lethal poison into Carol's water ($\neg AP$). Bob puts an antidote into Carol's water (BA). The water is lethal (L), if the poison is added without the addition of an antidote If Carol would consume the lethal water she would die (CD). Carol consumes her water (CC). Carol does not die ($\neg CD$).

We want to identify the cause of $\neg CD$. Intuition is again somewhat murky. However, it seems agreed upon that BA is not the cause of $\neg CD$. The uncertainty resides thus with $\neg AP$ and $\neg L$. Here we assume that neither of those should be considered a cause.

The modified Halpern and Pearl Definition

We define the binary causal model containing the equations

$$x_L := x_{AP} \wedge \neg x_{BA}$$
$$x_{CD} := x_{CC} \wedge x_L$$

with the context $\sigma := \{x_{AP} \mapsto \textbf{false}, x_{BA} \mapsto \textbf{true}, x_{CC} \mapsto \textbf{true}\}$.

We can observe that $\neg x_{AP}$ cannot be the cause of $\neg x_{CD}$, because x_{BA} holds thus $\neg x_L$ is fixed regardless of what variable we freeze and what intervention we perform on x_{AP}. Similarly, because $\neg x_{AP}$ holds, the same is true for x_{BA}. Moreover, because x_L does not hold, any form of intervention on x_{CC} cannot make x_{CD} true. However, if we intervene on the value of the variable x_L, setting it to true, we obtain x_{CD}. Hence, $\neg x_L$ can be considered a cause of x_{CD}. Lastly, notice that the equation $x_L := x_{AP} \wedge \neg x_{BA}$ represents in conjunction with the given context a case of Symmetric Overdetermination. That is, if we consider the conjunct $\neg x_{AP} \wedge x_{BA}$, then flipping the variables in question such that x_{AP} maps to true and x_{BA} maps to false we obtain that x_L and x_{CD} hold, i.e. $(\mathcal{I}, \sigma) \models [x_{AP}, \neg x_{BA}]x_{CS}$. Moreover, with AC1 being clearly satisfied and with AC3 checked above, we obtain that the conjunct is in fact a cause.

In summary, both the literal $\neg x_L$ and the formula $\neg x_{AP} \wedge x_{BA}$ are considered causes of x_{CD}. Moreover, the formalisation and the results can be found in [Hal15a].

A principled approach to actual causality

We define the binary causal model Δ containing the equations

$$x_L := x_{AP} \wedge \neg x_{BA}$$
$$x_{CD} := x_{CC} \wedge x_L$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

with the context $\sigma := \{x_{AP} \mapsto \textbf{false}, x_{BA} \mapsto \textbf{true}, x_{CC} \mapsto \textbf{true}\}$ and the timing such that x_{BA} comes after x_{AP}

$$\begin{aligned}\tau(\neg x_{AP}) &:= 1\\ \tau(\neg x_{L}) &:= 1\\ \tau(\neg x_{CD}) &:= 1\\ \tau(x_{BA}) &:= 3\\ \tau(x_{CC}) &:= 4\end{aligned}$$

given the satisfied literals under the model (Δ, σ). This timing is valid, particularly of note is the fact that $\tau(\neg x_L) = 1$, because the last chance at which x_L could have been satisfied was at time 1. Similarly, $\neg x_{AP}$ is sufficient for $\neg x_{CD}$, thus it happens exactly at that moment when the water is no longer lethal. That means it is determined that Carol will live before she drinks the water.

Clearly, $\neg x_{AP}$ does produce $\neg x_L$, because with $\neg x_{AP} \rightarrow \neg(x_{AP} \wedge \neg x_{BA})$ the set $\{\neg x_{AP}\}$ is sufficient. Moreover, $\neg x_L$ does produce $\neg x_{CD}$, because with $\neg x_L \rightarrow \neg(x_{CC} \wedge x_L)$ the set $\{\neg x_L\}$ is sufficient. Moreover, x_{AP} does not produce $\neg x_N$ in $(\Delta_{x_{AP}}, \sigma, \tau_{x_{AP}})$ because x_{AP} does not imply $\neg x_N$. Hence, any set sufficient for $\neg x_N$ remains sufficient once x_{AP} is removed. Therefore, $\neg x_{AP}$ is a cause of $\neg x_{CD}$. Furthermore, $\neg x_L$ is a cause of $\neg x_{CD}$, as no sufficient set of $\neg x_{CD}$ can contain x_L.

Looking at x_{BA} we can observe that given τ, x_{BA} cannot be a producer of $\neg x_N$. Therefore, it cannot be a cause of $\neg x_{CD}$.

Notice that if the timing would have been reversed x_{BA} and not $\neg x_{AP}$ would have been the cause. Moreover, if both $\neg x_{AP}$ and x_{BA} would have occurred simultaneously both would have been considered causes [BV18].

In summary, both literals $\neg x_{AP}$ and $\neg x_{AP}$ are considered causes of x_{CD}. Moreover, the formalisation and the results can be found in [BV18].

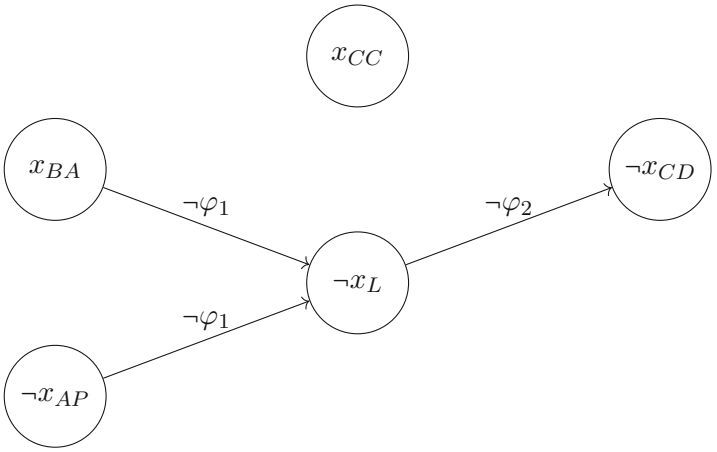
Possible Causal Process Semantic

We define the following causal theory

$$\begin{aligned}\varphi_1 &:= x_L \leftarrow x_{AP} \,||\, \neg x_{BA}\\ \varphi_2 &:= x_{CD} \leftarrow x_{CC}, x_L \,||\end{aligned}$$

Moreover, the situation suggests the world $\sigma := \{\neg x_{AP}, x_{BA}, \neg x_L, x_{CC}, \neg x_{CD}\}$. Now notice that φ_1 and φ_2 are blocked. Given this we can draw the causal process $\mathcal{I}$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN



The literals $\neg x_{AP}$ and $\neg x_L$ are both actual P-causes of $\neg x_{CD}$, due to both $(\neg x_{AP}, \neg x_L)$ and $(\neg x_L, \neg x_{CD})$ being non-trigger edges. By contrast, x_{BA} is not an actual P-cause of $\neg x_{CD}$, due to $(x_{BA}, \neg x_L)$ being a failure edge of a non-active causal mechanism.

In summary, both literals $\neg x_{AP}$ and $\neg x_{AP}$ are considered causes of x_{CD}. A more concise formalisation of this benchmark can be found in [DBV18]. Since the definitions differ slightly we derived the results independently.

Causal Inference

We define the following clausal causal theory

$$
\begin{aligned}
x_{AP} &\Rightarrow x_{AP} & \neg x_{AP} &\Rightarrow \neg x_{AP} \\
x_{BA} &\Rightarrow x_{BA} & \neg x_{BA} &\Rightarrow \neg x_{BA} \\
x_{CC} &\Rightarrow x_{CC} & \neg x_{CC} &\Rightarrow \neg x_{CC} \\
x_{AP}, \neg x_{BA} &\Rightarrow x_L & \neg x_{AP} &\Rightarrow \neg x_L \\
& & x_{BA} &\Rightarrow \neg x_L \\
x_{CC}, x_L &\Rightarrow x_{CD} & \neg x_{CC} &\Rightarrow \neg x_{CD} \\
& & \neg x_L &\Rightarrow \neg x_{CD}
\end{aligned}
$$

Moreover, the situation suggests the world $\sigma := \{\neg x_{AP}, x_{BA}, \neg x_L, x_{CC}, \neg x_{CD}\}$, thus the following causal rules are active.

$$
\begin{aligned}
x_{AP} &\Rightarrow x_{AP} \\
x_{BA} &\Rightarrow x_{BA} \\
x_{CC} &\Rightarrow x_{CC} \\
\neg x_{AP} &\Rightarrow \neg x_L \\
x_{BA} &\Rightarrow \neg x_L \\
\neg x_L &\Rightarrow \neg x_{CD}
\end{aligned}
$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Since we are able to derive $\neg x_{AP} \Rightarrow \neg x_{CD}$ and $x_{BA} \Rightarrow \neg x_{CD}$ using $\neg x_L \Rightarrow \neg x_{CD}$ while at the same time observing that $\mathbf{true} \not\Rightarrow x_{CD}$, the literals $\neg x_{AP}$, x_{BA} and $\neg x_L$ are considered causes.

This arises due to the symmetry that emerged from translating the causal model. A slight modification, i.e.

$$\begin{aligned} x_{AP}, \neg x_{BA} &\Rightarrow x_L & \neg x_{AP} &\Rightarrow \neg x_L \\ x_{CC}, x_L &\Rightarrow x_{CD} & \neg x_{CC} &\Rightarrow \neg x_{CD} \\ & & \neg x_L &\Rightarrow \neg x_{CD} \end{aligned}$$

Results in

$$\begin{aligned} \neg x_{AP} &\Rightarrow \neg x_L \\ \neg x_L &\Rightarrow \neg x_{CD} \end{aligned}$$

which means that only $\neg x_{AP}$ and $\neg x_L$ are causes of x_{CD}.

In summary, the literals $\neg x_L$, $\neg x_{AP}$ and x_{BA} are considered causes of x_{CD}. A more concise formalisation of this benchmark can be found in [Boc18a]. Hence, the result in our expanded version had to be derived independently.

4.2.7 Short Circuit

Benchmark 3.3.7 is an instance of "Short Circuit", which refers to the scenario, where an action is taken to prevent an inactive process, however, this triggers the process in the first place, which then has no effect because the original action prevents it from terminating. For a detailed discussion on this topic see Section 3.3.8.

This benchmark describes the following scenario.

> Carol is alive (CA). Alice puts a harmless antidote in Carol's water (AA). Adding antidote to the water, protects it against poison (WS - "water save"). If Alice puts the antidote into Carol's water, Bob will poison the water (BP) Adding poison to an unprotected water makes it toxic (WT). If Carol would drink toxic water she would die (i.e. inhibiting CS). Carol consumes her water and survives (CS).

Notice that the variable CA was only added to this benchmark, to ensure that CS is satisfied by default in the neuron diagram accompanying the example. Hence, it will be omitted. We want to identify the cause of CS. Here intuition dictates that either no event caused CS or possible $\neg WT$ and WS caused CS. However, as mentioned in Section 3.3.8, neither AA nor BP should be considered a cause.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The modified Halpern and Pearl Definition

We define the binary causal model containing the structural equations

$$x_{WS} := x_{AA}$$
$$x_{BP} := x_{AA}$$
$$x_{WT} := \neg x_{WS} \wedge x_{BP}$$
$$x_{CS} := \neg x_{WT}$$

with the context being $\sigma := \{x_{AA} \mapsto \textbf{true}\}$.

First, $\neg x_{WT}$ is a cause of x_{CS} in part because if we flip the value of x_{WT} to true, we obtain $\neg x_{CS}$. Second, x_{WS} is a cause of x_{CS} as well. That is, x_{WS} and x_{CS} both hold in the given model, the model obtained by flipping the value of x_{WS} results in $\neg x_{CS}$ and x_{WS} satisfies the minimality criteria. Third, x_{BP} is not a cause of x_{CS}. This is because intervening on x_{BP} does not impact the value of x_{WS} and thus x_{WT} remains to be false. Fourth, x_{AA} is a cause of x_{CS}, because if we intervene on x_{AA} and fix x_{BP} we obtain that x_{WT} evaluates to true, thus x_{CS} will evaluated to false.

In summary, the literals x_{AA}, x_{WS} and $\neg x_{WT}$ are considered causes of x_{CS}. Although the benchmark is discussed in [HH15], we chose to adhere to the formalisation presented in [Bau13]. Hence, the results had to be derived independently.

A principled approach to actual causality

We define the binary causal model containing the structural equations

$$x_{WS} := x_{AA}$$
$$x_{BP} := x_{AA}$$
$$x_{WT} := \neg x_{WS} \wedge x_{BP}$$
$$x_{CS} := \neg x_{WT}$$

with the context being $\sigma := \{x_{AA} \mapsto \textbf{true}\}$. As for the timing, we simply choose the constant timing 1.

First, we can observe that $\neg x_{WT}$ is a cause on x_{CS}. That is, the set $\{\neg x_{WT}\}$ is sufficient, the timing matches due to the timing function being constant, and both hold under the current model. By contrast, x_{WT} and x_{CS} cannot be both true, thus we fail to establish production in the negative case and demonstrate causation in the process.

Second, x_{WS} is a cause of x_{CS}. To establish this we demonstrate that x_{WS} produces $\neg x_{WT}$. First the set containing x_{WS} is sufficient because $x_{WS} \rightarrow \neg(\neg x_{WS} \wedge x_{BP})$ while the empty set does not. Moreover, there is no restriction w.r.t. to the timing function and x_{WS} and $\neg x_{WT}$ both hold in the model. Hence, we obtain x_{WS} causes x_{CS}. By contrast, $\neg x_{WS}$ cannot produce $\neg x_{WT}$, because if we intervene in the model we obtain x_{WT} due to $(\mathcal{I}_{\neg x_{WS}}, \sigma, \tau_{\neg x_{WS}}) \vDash x_{BP}$. Hence, $\neg x_{WS}$ cannot be an actual contributing

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

cause of $\neg x_{WT}$. With a similar argument, we can establish that x_{WS} is not an actual contributing cause for x_{CS}.

Third, x_{BP} is not a cause of x_{CS}. This arises due to the fact that it is not a producer of $\neg x_{WT}$, as no sufficient set for $\neg x_{WT}$ can contain x_{BP}. Hence, x_{BP} cannot be a producer of x_{CS} and thus no cause as well.

Fourth, x_{AA}. Clearly, x_{AA} is a producer of x_{WS} and therefore a producer of x_{CS}. By contrast, if we intervene to obtain the model $(\mathcal{I}_{\neg x_{AA}}, \sigma, \tau_{\neg x_{AA}})$ then we can observe that the set $\{x_{AA}\}$ is sufficient for $\neg x_{BP}$ and by extending the constant function $\tau_{\neg x_{AA}}$ to all literals, we obtain that $\neg x_{AA}$ produces $\neg x_{BP}$. In an analogue to x_{WS} in the original model, $\neg x_{BP}$ subsequently produces $\neg x_{WT}$, which as we know produces x_{CS}. Hence, we can conclude that $\neg x_{AA}$ produces x_{CS} and therefore fails to be a cause of x_{CS}.

In summary, nothing caused x_{CS}. Moreover, a slightly different formalisation and the respective results can be found [BV18].

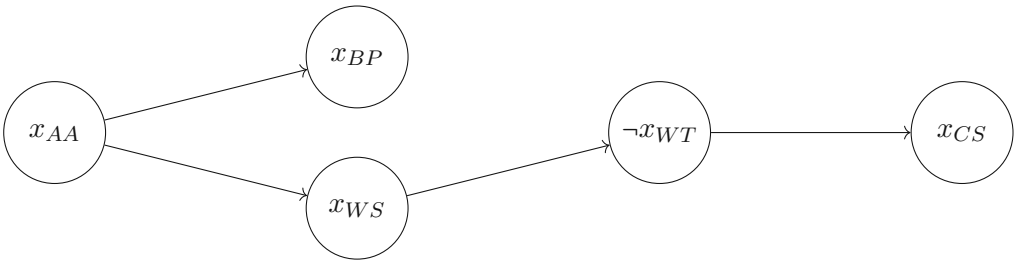
Possible Causal Process Semantic

We define the following causal theory

$$\begin{aligned}
\varphi_1 &:= x_{WS} \leftarrow x_{AA} \parallel \\
\varphi_2 &:= x_{BP} \leftarrow x_{AA} \parallel \\
\varphi_3 &:= x_{WT} \leftarrow x_{BP} \parallel \neg x_{WS} \\
\varphi_4 &:= \neg x_{CS} \leftarrow x_{WT} \parallel
\end{aligned}$$

Moreover, the world $\sigma := \{x_{AA}, x_{WS}, x_{BP}, \neg x_{WT}, x_{CS}\}$ is suggested. Observe that both φ_1, φ_2 are active, applicable and satisfied, that φ_3 is both active and failed because of x_{WS}, and that φ_4 is simply blocked due to $\neg x_{WT}$.

Given this we can draw the causal process $\mathcal{I}$



The edges (x_{AA}, x_{WS}) and (x_{AA}, x_{BP}) are both trigger edges. The edge $(x_{WS}, \neg x_{WT})$ is a failure edge and the edge $(\neg x_{WT}, x_{CS})$ is a non-trigger edge. Now given that φ_3 is active we can find the path from x_{AA} over x_{WS} to x_{CS} that consists only of trigger, non-trigger and failure edges of active causal mechanism. Hence, we can conclude that

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

x_{AA}, x_{WS}, $\neg x_{WT}$ all are causes of x_{CS}. By contrast, there does not exists a path from x_{BP} to x_{WS} and thus x_{BP} cannot be a cause.

In summary, both literals x_{AA}, x_{WS} and $\neg x_{WT}$ are considered causes of x_{CD}. No corresponding formalisation was found. Hence, both the model and the results were produced independently of the literature.

Causal Inference

We define the following clausal causal theory

$$
\begin{aligned}
x_{AA} &\Rightarrow x_{AA} & \neg x_{AA} &\Rightarrow \neg x_{AA} \\
x_{AA} &\Rightarrow x_{WS} & \neg x_{AA} &\Rightarrow \neg x_{WS} \\
x_{AA} &\Rightarrow x_{BP} & \neg x_{AA} &\Rightarrow \neg x_{BP} \\
\neg x_{WS}, x_{BP} &\Rightarrow x_{WT} & x_{WS} &\Rightarrow \neg x_{WT} \\
& & \neg x_{BP} &\Rightarrow \neg x_{WT} \\
\neg x_{WT} &\Rightarrow x_{CS} & x_{WT} &\Rightarrow \neg x_{CS}
\end{aligned}
$$

Moreover, the situation suggests the world $\sigma := \{x_{AA}, x_{WS}, x_{BP}, \neg x_{WT}, x_{CS}\}$, thus the following causal rules are active.

$$
\begin{aligned}
x_{AA} &\Rightarrow x_{AA} \\
x_{AA} &\Rightarrow x_{WS} \\
x_{AA} &\Rightarrow x_{BP} \\
x_{WS} &\Rightarrow \neg x_{WT} \\
\neg x_{WT} &\Rightarrow x_{CS}
\end{aligned}
$$

Given this sub-theory, we obtain that x_{AA}, x_{WS}, $\neg x_{WT}$ are all causes of x_{CS} due to transitivity. Moreover, we cannot infer that x_{BP} is a cause of x_{CS}.

In summary, both literals x_{AA}, x_{WS} and $\neg x_{WT}$ are considered causes of x_{CD}. No corresponding formalisation was found. Hence, both the model and the results were produced independently of the literature.

4.2.8 Comparison

Here we briefly summarise the results obtained in the previous sections. To that end, Table 4.1 provides a concise summary, by listing the causes obtained from applying the definitions introduced in Section 4.1 to the considered benchmarks (see Section 3.3).

Among the results for Benchmark 3.3.1 the definition `HP-15` is particularly interesting. Indeed, considering that the other definition can only identify causes if they are encoded as literals, `HP-15` permits a more flexible notion of cause. That is, given the sketched situation, `HP-15` identified that due to the overdetermination we cannot isolate singular

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

causes for the shattering of the window. Hence, rather than considering each AF and BF as causes individually, it is their conjunct that ensures WS.

For the scenario sketched in Benchmark 3.3.2, both `BV-CM` and `PCPS` comply with intuition, because both recognise that the choice of track is immaterial for the arrival of the train. The definition `BV-CM` accomplishes this by demanding causes to be asymmetric w.r.t. production. Meaning that for AF to be a cause, only AF and not $\neg AF$ can produce TD. The definition `PCPS` achieves this by distinguishing between two types of events: triggering and enabling conditions. The flicking of the switch only enables the train to travel on track A, i.e. the flicking of the switch only establishes the background conditions, which given this definition cannot be causes. However, it may be unclear what a triggering and what an enabling condition should be, as AF could be considered as a triggering event as well. To remove this uncertainty we followed the model structure presented in [DBV19]. By contrast, `HP-15` and `BCI` both declare the flicking of the switch a cause. As discussed in Section 3.3.3, [HH11] suggests that the discrepancy between intuition and the inference of his definition is a result of the implicit assumption that one of the tracks could be blocked. Hence, he argues that a model that captures this assumption may be better suited for this situation, e.g. such a model would contain the equation $TD := (AF \wedge \neg BA) \vee (\neg AF \wedge \neg BB)$, where BA and BB represent whether the respective track is blocked or not. Given this reformulation, `HP-15` complies with the stated intuition. However, notice that in this reformulation AF is part of the equation for TD. Hence, one reason why AF suddenly ceases to be a cause in the modified model is that we eliminate the hypothetical scenarios where the track travels on no or on both tracks at the same time. That is, adding TA and TB as auxiliary variables creates possible worlds which are immediately discounted by humans. Thereby, creating a rift between intuition and formal inference. For example, if we select a causal model with the structural equations of $TD := AF \vee \neg AF$ then clearly no intervention on AF can result in the train not arriving. In this simplistic model, both `HP-15` and `BCI` do not recognise TD as cause.

The results for Benchmark 3.3.3 are uniform in the sense that all definitions agree with the intuitively correct answer. Particularly of note is the `BV-CM`. Here the inclusion of a timing function allows for an elegant encoding of the problem. That is, the causal model can have the same structure as in the Symmetric Overdetermination scenario. However, rather than both events acting at the same time, the selected timing ensures that one event occurs after the effect, breaking the overdetermination. As for `HP-15`, `BCI` and `PCPS` this behaviour is enforced by introducing auxiliary variables that mimic the progression of time. However, given the observation that adding auxiliary variables can lead to undesired inferences, as in the Switch scenario of Benchmark 3.3.2, this may not be the ideal choice of action. `PCPS` differs only by declaring AH to be an enabling condition of BH, thus ensuring that the process set in motion by BF can only terminate if the process set in motion by AF fails.

On the Benchmark 3.3.4 only the definition `BV-CM` disagrees with the supposedly correct answer. That is, rather than declaring AF to be the cause of WS, the window shatters

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

without any cause. While initially counter-intuitive this choice was deliberate, as in their view Early Preemption is closely related to Switch. That is, irrespective of the value of AF the window will always shatter. Hence, if we would like to strengthen this resemblance, we could add an auxiliary variable between AF and WS, which given the results obtained for Switch would be considered a cause of WS. By contrast, all of the definitions agree that BF cannot be a cause of WS.

As for Benchmark 3.3.5 only the definition PCPS fails to capture the stated intuition. This is primarily due to the choice of modelling BSA as an enabling condition for AF. This eliminates the causal influence of CSB and CI. The most important result to highlight for this benchmark is that no definition declared BI to be a cause of WS.

For Benchmark 3.3.6 the results are slightly more diverse. Most importantly, all definitions declare the non-lethality of the water to be a cause of Carol's survival. Although, not entirely in line with intuition, void of any form of normality assumptions declaring $\neg L$ to be a cause on $\neg CD$ seems to be a reasonable inference. However, as soon as we recognise that adding the assumption "$\neg L$ is the normal state of the world", this inference becomes less clear. The definition HP-15 adds to this by declaring the conjunct $\neg AP \wedge BA$ to be a cause. However, it does not deem either of the individual literals as a cause, because to the part of the causal model that establishes lethality, is similar to an instance of Symmetric Overdetermination. Moreover, the similarity to Symmetric Overdetermination also explains the behaviour of BCI, which in addition to $\neg AP$ also declares BA as a cause of $\neg CD$. The scenario slightly shifts when considering BV-CM, which declares $\neg AP$ to be a cause. That is, while from a causal model perspective we are presented with a case of Symmetric Overdetermination, the temporal component in the story indicates that $\neg AP$ comes before BA. Hence, we are in fact faced with a case of Late Preemption, which can be captured using the timing function without modifying the underlying causal model. By contrast, PCPS behaves quite differently from Symmetric Overdetermination. In fact, the behaviour resembles a complete inverse of the situation. That is, in Symmetric Overdetermination all paths are trigger paths indicating that all causal mechanisms terminated successfully. Whereas in this case all paths are either failure or non-trigger paths, which indicates that all causal mechanisms either failed or were not initiated in the first place.

On the Benchmark 3.3.7 all definitions agree with the non-controversial intuition that BP is not a cause of CS. However, all definitions except BV-CM declare AA a cause of CS as well. However, given the fact that AA essentially generates its own relevance, such an inference seems counter-intuitive. One reason for this particular difference between the formalisms is that the part of the scenario that determines the toxicity of the water resembles to some extent a Switch scenario. Hence, it is no surprise that the definitions HP-15 and BCI which declared the flicking of the switch to be a cause of the train arrival in Benchmark 3.3.2, also declare AA to be a cause of CS. A slightly different picture arises in the case of PCPS, because of a discrepancy in behaviour between this and the Switch scenario. The reason behind this is that in this case the switch event, i.e. AA actually takes on the role of a triggering condition. That is adding the antidote triggers

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the water to be save against poison and triggers Bob to add poison to the water. By contrast, if we would rephrase this to "adding the antidote enables the water to be save against poison" and "adding the antidote enables Bob to add poison to the water", then AA would seize to be a cause. As already mentioned the definition BV−CM manages to avoid this inference by requiring that causes are asymmetric and since $\neg AA$ also produces CS it denies AA the status of cause.

To summarise, the Benchmark 3.3.1 nicely highlights a major advantage of HP−15, namely that its causal attribution extends beyond mere literals. Emphasising, that both AF and BF individually are only part of the cause of the WS. Not because they are not able to produce WS on their own, but because both events have to fail in order to prevent the window from shattering. One major limitation of the sketched comparison is the fact that we have limited ourselves to binary scenarios only. That is, the selected benchmarks could all be encoded using deterministic binary variables. Hence, one major advantage of HP−15 remained hidden. Namely, it is defined to cope with multi-valued variable values and is therefore able to capture scenarios that go beyond the capabilities of the other definitions. Especially the ability to deal with probabilities in a rudimentary manner seems to be a desirable feature. From our point of view the most troublesome behaviour of this definitions occurs in the case of Switch (Benchmark 3.3.2). That is, we think that the flicking of the switch is incorrectly classified as cause. In particular, the definition is not robust to the addition of the auxiliary variables x_{TA} and x_{TB}, because in the reduced model containing only the structural equation $x_{TA} := x_{AF} \vee \neg x_{AF}$ the "correct" inference is made. This is due to the fact that in the extended model we can produce a hypothetical scenario that results in the train not arriving. The feature missing here is a method of pruning "impossible" hypothetical scenarios, which would make the definition more robust to the addition of auxiliary variables.

As seen in the Benchmark 3.3.2, the definition BV−CM captures a particularly desirable property of causation in an elegant fashion, namely asymmetry. That is, because both the flicking of the switch and the omission of flicking the switch results in the train arriving. Hence, the action taken is immaterial for the outcome and should therefore not be considered a cause. This is precisely the reason why BV−CM in unable to identify any cause for the shattering of the window in the Early Preemption scenario. Put bluntly, in such cases the outcome is already predetermined and thus given the required asymmetry no event can actually cause it. Moreover, we can observe the benefits of BV−CM's timing function in two separate examples, i.e. the Benchmark 3.3.3 and the Benchmark 3.3.4. That is, while both scenarios have proven to be challenging for past definitions, BV−CM demonstrates that this challenge can be circumvented by simply extending causal models by a temporal dimension that goes beyond the addition of auxiliary variables. This allows not only for more elegant and robust modelling of such situations, but provides additional insights. For example, as seen in Benchmark 3.3.6, we could observe that a valid timing for the discussed scenario requires that Carol is guaranteed to survive at the very moment at which Alice refrains from adding poison to the water. Hence, demonstrating in a painfully explicit manner that adding the antidote or drinking the water is immaterial for

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

the survival of Carol. Compared to HP−15, BV−CM has the disadvantage that causes are restricted to literals. Hence, in the context $\{x_{AF} \mapsto \mathbf{t}, x_{BF} \mapsto \mathbf{t}\}$we can not distinguish between the causal model containing the structural equation $x_{WS} \coloneqq x_{AF} \vee x_{BF}$ and the one containing the equation $x_{WS} \coloneqq x_{AF} \wedge x_{BF}$ on the declared causes alone, as in both cases x_{AF} and x_{BF} are considered causes.

The chosen examples were not ideal to highlight the unique strength of PCPS, namely the distinction between enabling and trigger conditions. To demonstrate the utility of this feature consider the following scenario.

> A forest fire (FF) was ignited by a spark (S) due to the presence of dry grass (D) and oxygen (O). What caused the forest fire?

By modelling this situation using the causal mechanism $x_{FF} \leftarrow x_S \parallel x_D, x_O$, we can satisfy the intuition that the spark was the cause of FF, while both D and O were merely background conditions. By contrast, with HP−15 the straightforward causal model with the structural equation $x_{FF} \coloneqq x_S \wedge x_D \wedge x_O$ we would declare all literals individually as cause of $x_F F$. A similar picture arises for BV−CM and BCI as well. However, a deficit of PCPS is the fact that it requires that variables can change their value only once. For example, we are not able to model a light switch, i.e. flicking the switch once turns the light on and flicking it again turns the light off. Hence, it is missing some mechanism that allows for a more liberal definition of a causal theory [DBV18]. Moreover, as with BV−CM, the restriction to literals as causes makes the disjunctive scenario, i.e. $\{x_{WS} \leftarrow x_{AF},\ x_{WS} \leftarrow x_{BF}\}$, and the conjunctive scenario, i.e. $\{x_{WS} \leftarrow x_{AF}, x_{BF}\}$, indistinguishable if both x_{AF} and x_{BF} hold.

Lastly, the definition BCI performs similar to HP−15, on this particular set of examples, which seems to be intentional, as BCI is an attempt to provide a regularity based perspective to the recent advances made by the counterfactual tradition of causality. This, unfortunately, makes it difficult to pinpoint scenarios where BCI provides a unique perspective to causality. However, what sets BCI apart it allows for causal inference using a proper logical foundation [Boc18a]. Due to its similarity with HP−15, the criticism about the Switch scenario in Benchmark 3.3.2 applies here as well. That is, the literal x_{AF} is considered a cause, even if the outcome is predetermined. Moreover, in the simplified version of this scenario, i.e. the one without auxiliary variables, x_{AF} would still be the cause of x_{TA}. To be precise, due to the fact that disjuncts are not allowed the causal theory in this scenario would be $\{x_{AF} \Rightarrow x_{TA}, \neg x_{AF} \Rightarrow x_{TA}\}$ and therefore just $\{x_{AF} \Rightarrow x_{TA}\}$ in the case where x_{AF} holds. Furthermore, similar to BV−CM and PCPS, BCI can not distinguish between the conjunctive and disjunctive scenario.

Due to the choice of benchmarks there are some severe limitations with the above comparison. That is, the presented benchmarks can be captured on a propositional level and ask only for binary causal attribution, i.e. something is a cause or not. Hence, the investigations miss that none of the discussed formalism can capture first-order statements and that apart from HP−15 none of them is capable of expressing causality in

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

a quantitative fashion. Both avenues are explored in the causality literature. For example, [BS18] uses situation calculus to move towards a first-order definition of token causality and [Hal16a] discusses the connection between causes and degrees of responsibility using `HP-15`. However, overall those two features are under-represented in the discussion.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Scenario	Example	Intuition	HP-15	BV-CM	PCPS	BCI
Symmetric Overdet.	3.3.1	AF, BF (or $AF \wedge BF$)	$AF \wedge BF$	✓	✓	✓
Switch	3.3.2	TA	AF, TA	✓	✓	AF, TA
Late Preemption	3.3.3	AF, AH	✓	✓	✓	✓
Early Preemption	3.3.4	AF	✓	∅	✓	✓
Double Preemption	3.3.5	AI, AF, CI, CSB, $\neg BSA$	✓	✓	AI, AF	✓
Bogus Preemption	3.3.6	∅ (or $\neg L$)	$\neg L$, $\neg AP \wedge BA$	$\neg AP$, $\neg L$	$\neg AP$, $\neg L$	$\neg L$,$\neg AP$, BA
Short Circuit	3.3.7	∅ (or WS, $\neg WT$)	AA, WS, $\neg WT$	✓	AA, WS, $\neg WT$	AA, WS, $\neg WT$

Table 4.1: This table summarises the results obtained by applying the selected formalisms to the Benchmarks in Section 3.3. Each cell contains the literals (or formulas) that are deemed causes by the respective definition. To enhance readability the check mark indicate that the particular definition complies with intuitively "correct" answer for the scenario captured in the benchmark. The cases where no cause is present are indicated using ∅.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Conclusion

This thesis provided a systematic review of the causality literature on three different levels of granularity. In Chapter 2 we studied the structure of a large subsection of the causality literature to identify important authors, publications and research communities. Building on Chapter 2, Chapter 3 surveys the set of important publications to identify important formal languages, definitions and benchmarks used in the causality literature. Chapter 4 further narrows the scope, by testing the capability of some important definitions from Chapter 3 against some of the introduced benchmarks.

To identify important publications, authors and research communities working in the field of token causality, we surveyed approximately 5000 publications. That is, we started by collecting all publications in "Journal Knowledge-Bases Systems", the "Journal Artificial Intelligence", the "Journal Artificial Intelligence and Law" and "International Joint Conferences on Artificial Intelligence Organization" that were published between 01.2017 and 3.2020. Using a simple key-word search we reduced 4223 unique publications, to a manageable set of 37 publications. After employing several forward-snowball, backward-snowball and filter steps we obtained a total of 872 publications. Those publications were subject to closer inspection and further filtering which provided us with 294 relevant publications, which were further reduced to 107 publications by considering only those that had been published in the past decade.

Using the collected publication we constructed several graphs. The two most important ones were the publication graph $\mathcal{G}_p$ and the merged graph $\mathcal{G}_m$. The former graph, was obtained by extracting the citation relation from the bibliographies of each of the 107 papers, this allowed us to employ centrality measures to identify important publications. Using the rankings induced by those centrality measures we were able to identify 36 important publications which were subjected to further study in subsequent chapters. Notable mentions are [Wes15], [BS17], [HH11], [GDG+10] and [HH15], all of which are ranked highly across all measures. The fact that all those publications use causal models as their preferred method of encoding causal relations hints at the dominance of this

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

framework throughout the causality literature. A claim that is further supported by the observations made in Chapter 3. The graph $\mathcal{G}_m$, was obtained by both connecting the authors based on the citation relation of $\mathcal{G}_p$ and by connecting them based on the co-authorship relation. Here, the most important findings were that the authors Lifschitz, Icard, Bochman, Eberhardt, Hitchcock, Gerstenberg, Lagnado and Halpern consistently score high across each ranking. Furthermore, we were able to observe that there have been collaborations between Bochman and Lifschitz; Halpern and Hitchcock; Gerstenberg and Icard. The first two were particularly important for this thesis. That is, both Lifschitz and Bochman focus on variants of the causal theory put forward in [MT+97] and tend to approach causality from a regularity theoretic point of view. By contrast, Halpern and Hitchcock strongly adhere to the structural equation framework. Their investigations into causality, while emerging from the counterfactual tradition, recently incorporate some regularity theoretic tools, e.g. extending causal models with normality rankings. Some auxiliary investigations revealed that with a total of 114 publications the decade between 2000 and 2010 was the most productive one. However, as we mostly neglected the part of the literature concerned machine learning, it seems reasonable to assume that there is actually an increase in publications discussing causality over time. However, this hypothesis was not tested. Nevertheless, notable publications of this decade are "Nonmonotonic Causal Theories" [GLL+04], "Causes and Norms" [HK09], "Prevention, Preemption, and the Principle of Sufficient Reason" [Hit07a], "Two Concepts of Causation" [Hal04] and "Structural Equations and Causation" [Hal07].

Analysing these 36 important publications, we could identify 18 unique formal languages used for encoding causal relationships, 32 unique token causality definitions and more than 20 benchmarks used for testing said definitions. Moreover, we tracked how often each of those constructs were referenced within the given set of publications in order to gauge their popularity.

By far the most discussed language family is the one building on causal models, with the CP-Logic (causal and probabilistic Logic) family and the non-monotonic causal theory tying for a distant second place. The causal model family contains the greatest amount of languages, making it by far the most developed strain of formalisms. Introduced by Pearl, it assumes that causal mechanisms governing the world can be described by a set of random variables and a set of deterministic structural equations. This allows one to condense all type-causal relations that may influence a variable into a single, asymmetric equation. The CP-Logic family is closely related to logic programming. Among others, it contains two rather distinct members, the first heavily emphasising the use of probabilities in encoding causal relations while the second taking a more process-orientated view by encoding causal relations as causal mechanisms that have both triggering and enabling conditions. Non-monotonic causal theory simply extends an ordinary propositional language with a causal relation, which expresses that a proposition causes another proposition. This inference relation is a more restrictive variant of the production inference relation, a defining feature of which is its failure to satisfy the reflexivity postulate.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Among the plethora of token causality definitions, by far the most discussed are the three definitions put forward by Halpern (and Pearl), which naturally are defined in terms of causal models. Moreover, given the fact that more than 12 additional definitions were build on the causal model framework, it is quite apparent that the ideas of Halpern and Pearl still influence the causality literature heavily. By contrast, although there are various interesting definitions that utilise other languages, e.g. situation calculus, non-monotonic causal theories, CP-Logic and others, their investigations rarely extend beyond their publication of origin. However, it is possible to detect two trends among the definitions. Firstly, the boundaries between the traditions start to blur. For example, there are counterfactual definitions that are incorporating regularity theoretic ideas such as normality into their definitions, e.g. [Hal08] and [Wes15]. Moreover, even the definition from [Boc18a], which sees itself firmly rooted in the regularity theoretic tradition, has some counterfactual flavours. Secondly, among the newer definitions there seems to be a greater emphasis on time and processes. For example, [BV18] extends the causal model framework using a timing function, [DBV19] models causal relationships as processes and most importantly the significant number of new definitions are using situation calculus, i.e. [BS18], [LBV19] and [KS20].

In Chapter 4 we highlighted the causal model based definition found in [BV18], the CP-Logic based Possible Causal Process approach found in [DBV19] and the Non-Monotonic Theory based approach found in [Boc18a], because they are the most recent definitions from each of the three most popular language families. Moreover, we further discussed the newest incarnation of the definitions put forward by Halpern, namely the modified Halpern and Pearl definition introduced in [Hal15a]. However, given the many definitions detected, this selection only provides a small glimpse into the techniques used for formalising causality. For example, the definitions utilising situation calculus could provide additional insight into causation, because as of now they belong to the few definitions that actually tackled causality from a first-order perspective. Furthermore, we could observe that most definitions are grounded in either the counterfactual or the regularity theoretic tradition. However, there are some definitions that take ideas from other philosophical traditions, e.g. the probabilistic interpretation of Halpern's definition put forward by [FG17].

The primary objective of token causality literature seems to be the development of a definition that corresponds to the intuitive human understanding of causation. To demonstrate or refute the claim that a definition satisfies the specified goal, the literature accumulated a significant number of benchmarks, which represent edge cases that have proven to be troublesome to capture. By far the most commonly used ones are: *Symmetric Overdetermination*, which refers to the scenario where multiple processes, all of which producing the same outcome, terminate at the same time; *Switching*, which refers to the scenario where there exists an event that triggers one of two processes both of which have the same outcome, thus making the event immaterial for the outcome of the scenario; *Late Preemption*, which refers to the scenario where there are two causal processes running in parallel, both would produce the same outcome, but one process terminates before the

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

other does, thus bringing forth the outcome and rendering the second process irrelevant; *Early Preemption*, which refers to the scenario where there are two causal processes, both would produce the same outcome, but one process terminates before the other can even start; *Double Preemption*, which refers to the scenario where a process that would have prevented another process, was prevented by an entirely different process itself; *Bogus Preemption*, which refers to the scenario where an action is taken to interrupt an inactive process; *Short Circuit*, which refers to the scenario where action is taken to prevent an inactive process, however, this in fact triggers the process in the first place, which then has no effect because the original action prevents it from terminating.

Considering the listed benchmarks, there are two significant problems. The first is that for some it might not be clear what the correct answer should be. The second is that it is often not clear how to formalise the scenarios sketched in those benchmarks, which is not ideal given that many of the presented definitions are highly sensitive to slight changes in a model, e.g. adding a simple auxiliary variable may change the resulting inferences drastically. Hence, debates surrounding the correct definition of causality, are tainted by disagreements on the correct answers and even the correct formalisation, thus prohibiting a clean comparison between various formalisms.

Nevertheless, those benchmarks are used in Chapter 4 to compare the four definitions. While we were able to find several evaluations of those benchmarks in the literature, this was not always the case. Hence we were sometimes required to formalise and evaluate those benchmarks independently of the literature to properly compare the definitions. While most definitions comply more or less with what humans would intuitively consider causes, it is still possible to highlight some differences. The definition put forward by Halpern, provides a relatively simple inference procedure that extends the notion of causes to entire conjuncts of literals. By contrast, all the other definitions restrict themselves to literals only. This results in an arguably more intuitive inference on the Symmetric Overdetermination scenario. By contrast, it provides a less desirable result on the Switch scenario, which in our opinion is best handled by the causal model based approach developed by Beckers, which encodes the assumption that an event can only be a cause of another event, if the absence of the event does not result in the same event. This approach further differentiates itself from Halpern's definition by explicitly including time in the inference procedure, allowing for an elegant handling of the time-dependent scenarios, e.g. Late Preemption or Early Preemption. A refreshingly new perspective on causation is brought forth by Denecker, in the form of the Possible Causal Process Semantic. The unique feature of this approach is the distinction between enabling and trigger conditions. Although making modelling slightly more difficult, it elegantly addresses the issue of background conditions and whether they should be considered causes or not. Lastly, what sets Bochman's definition apart is that it provides a proper logical foundation for performing causal inference.

From our point of view, the two most interesting issues surrounding token causality are time and normality. First, while type causal relationships can be atemporal, e.g. temperature and altitude, establishing token causality always requires some form of

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

temporal progression, i.e. a cause cannot occur before its effect. However, many formalisms leave the progression of time more or less implicit. Additionally, paying more attention to time would resolve the need for cyclic models, e.g. feedback loops could be unravelled along the temporal axis. Second, normality or more generally additional context. The idea of incorporating normality is a result of the observation that humans declare an event to be a cause, if it is something that is out of the ordinary. For example, for a fire we require oxygen and a spark. If we are on earth the spark would clearly be the cause of the fire, because oxygen is assumed as a given. By contrast, if we would be in space, then the assertion that both should be considered a cause becomes more reasonable. It is postulated that appealing to normality may address such concerns. Another important avenue that should be explored in greater detail, is the idea of soft causation. That is, while some definitions recognise that not all causes contribute equally to the occurrence of the effect, the literature is starved of investigations on how to quantify the extent to which an event contributed to the occurrence of an event.

However, above all this it is hard to disagree with [GDG+10], and ask whether the inductive "example first" approach is actually fruitful. Especially if we want to formalise more complex scenarios, that are intertwined with an individual understanding of the world. For example, if we want to identify what caused the bad grades of a student, was it the teacher, was it the friend circle, the socio-economic status of the parents, the race of the student, the fact that his dog died last year, and so on. In such complex scenarios, it already seems impossible to formulate a sufficiently inclusive model of the word to capture all those different factors. Moreover, if we want to assess the causes using counterfactuals, we would actually require data covering all the hypothetical scenarios such that we could precisely assess their influence on grades. Hence, even if we would have a correct definition of what a token cause is, such form of inference becomes quickly infeasible due to the immense amount of data required for establishing the connection between the variables.

However, it is important to realise that this concern does not diminish the utility of those definitions in more controlled settings, e.g. physical laws, game theory, computer programs, and so on. In a similar fashion, this survey can conclude, that there does not seem to exist a definition that perfectly captures token causality. Indeed, given the vagueness of human intuition there may never be one definition that is beyond criticism, however, the modern definitions of token causality seem to be sufficiently robust to provide ample utility in a variety of settings.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

Appendix

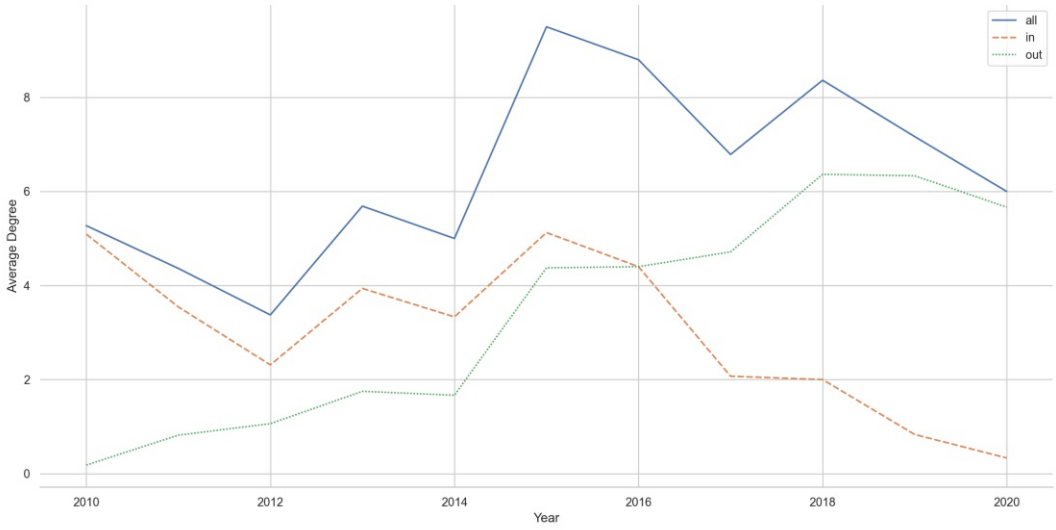


Figure 6.1: A line graph depicting the average in-degree, out-degree and overall degree of the publications in $\mathcal{G}_p$.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Figure 6.2: A subgraph of $\mathcal{G}_m$, depicting Group 1.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

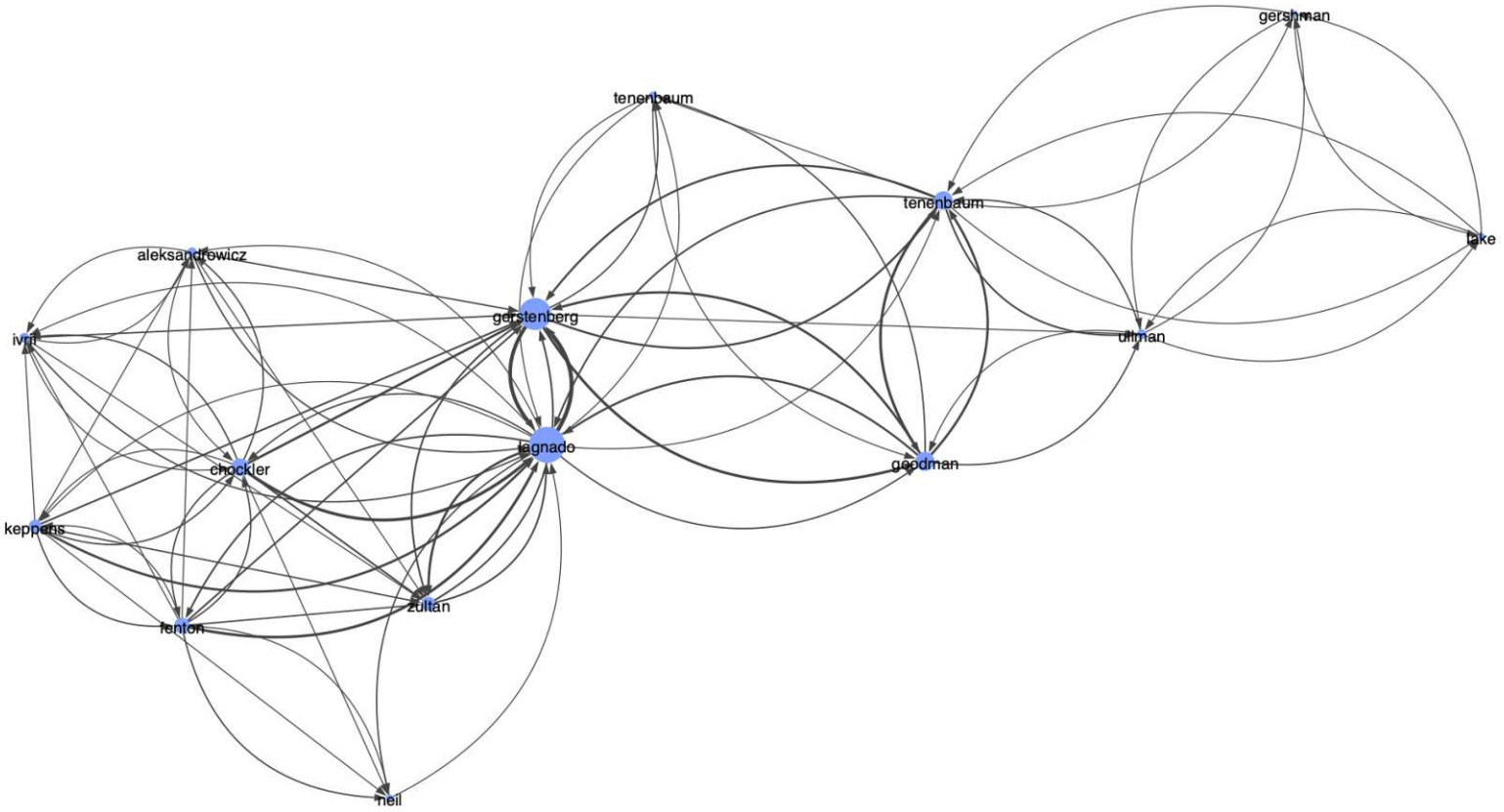


Figure 6.3: A subgraph of $\mathcal{G}_m$, depicting Group 2.

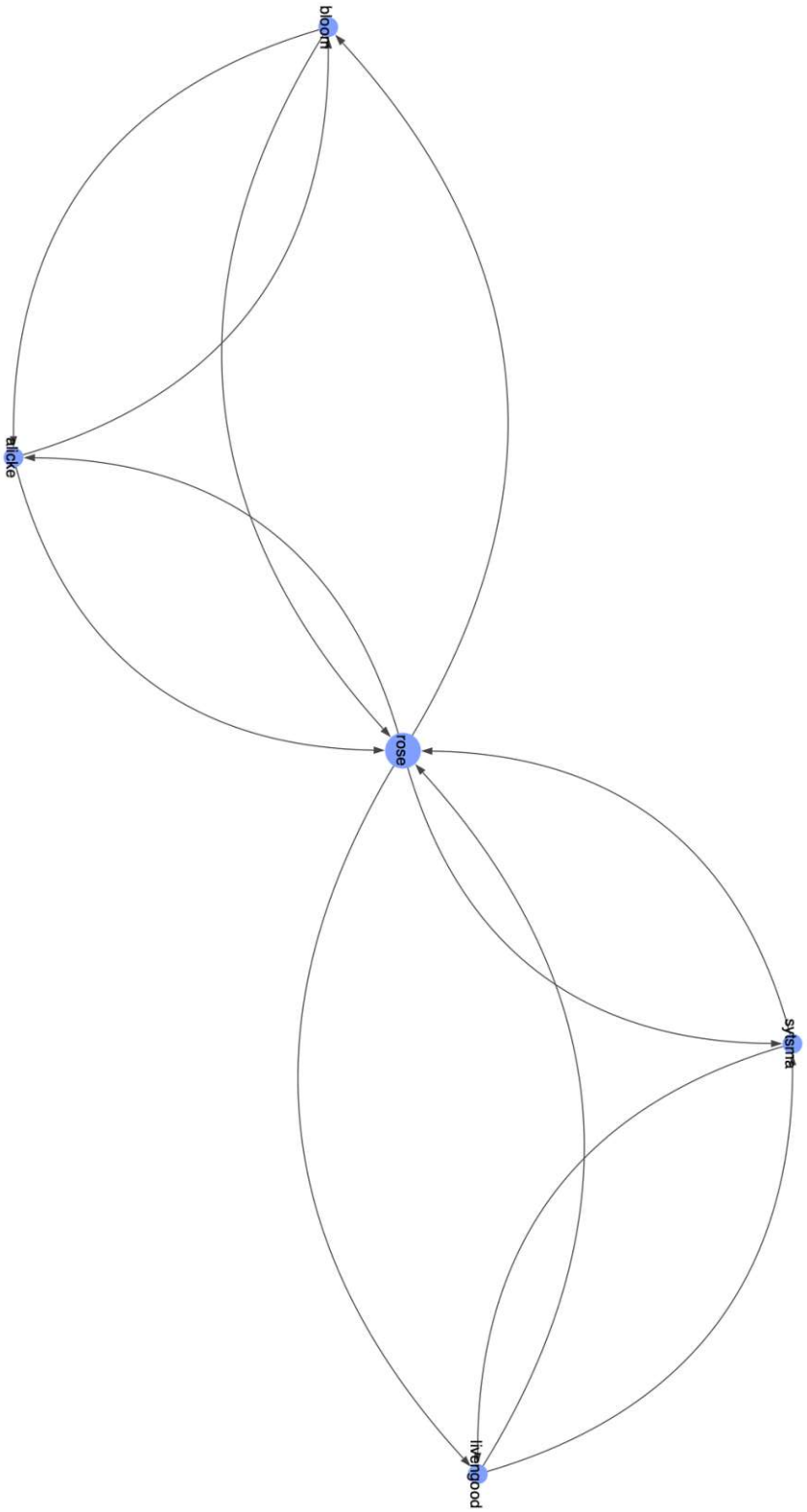


Figure 6.4: A subgraph of $\mathcal{G}_m$, depicting Group 3.

TU Bibliothek Your knowledge hub WIEN
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

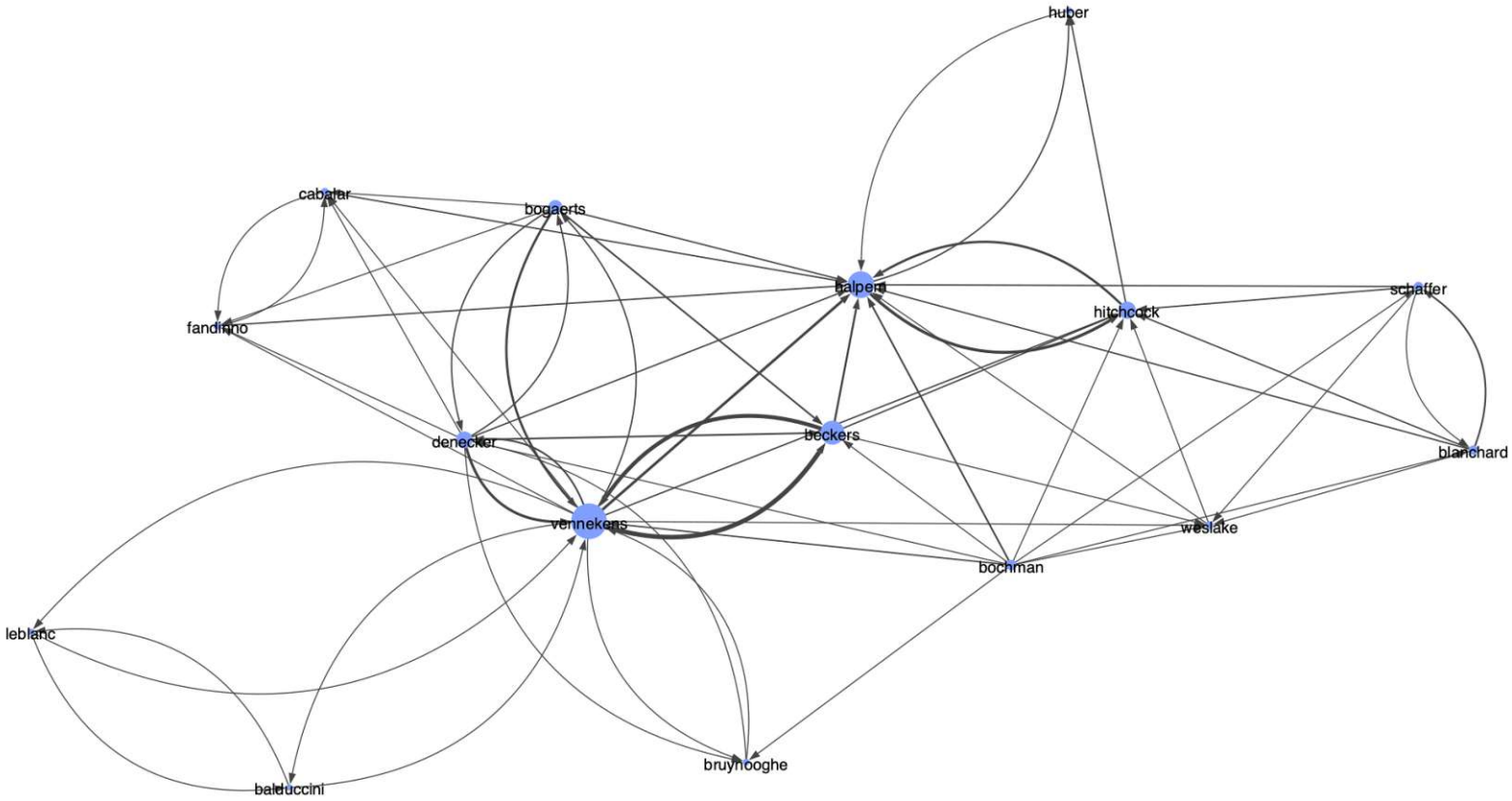


Figure 6.5: A subgraph of $\mathcal{G}_m$, depicting Group 4.

Figure 6.6: A subgraph of $\mathcal{G}_m$, depicting Group 5.

TU Bibliothek Your knowledge hub WIEN
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

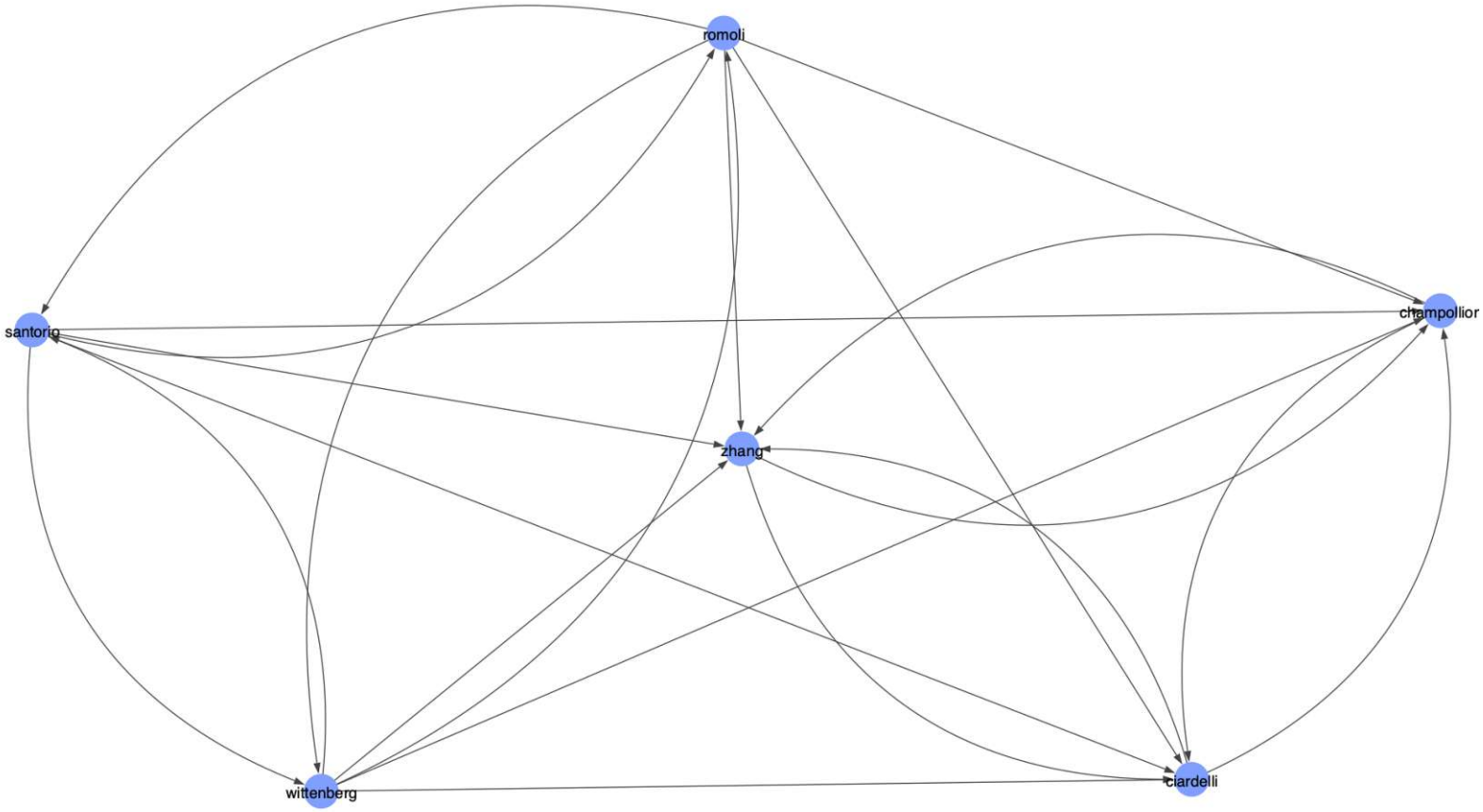


Figure 6.7: A subgraph of $\mathcal{G}_m$, depicting Group 6.

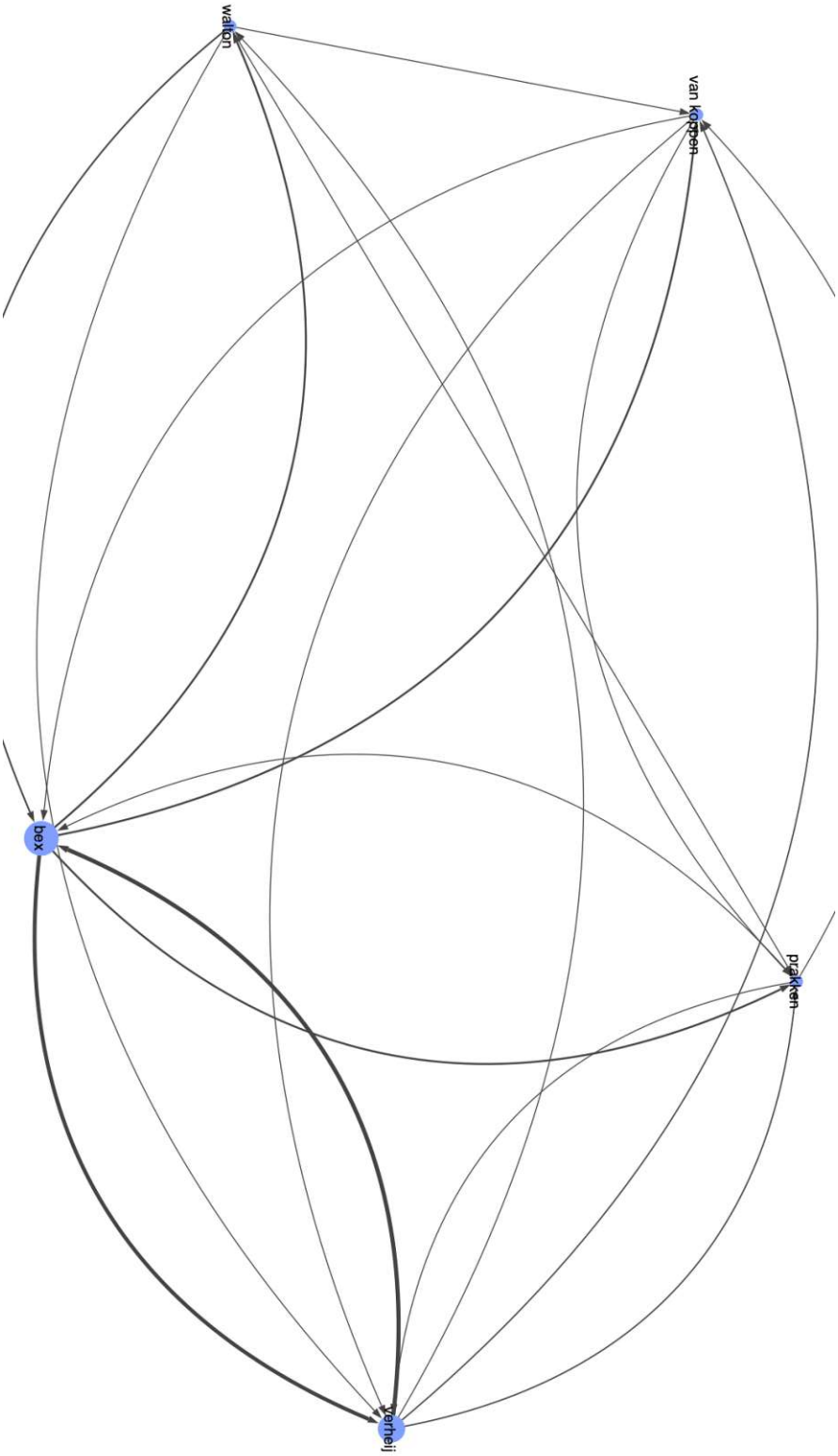


Figure 6.8: A subgraph of $\mathcal{G}_m$, depicting Group 7.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

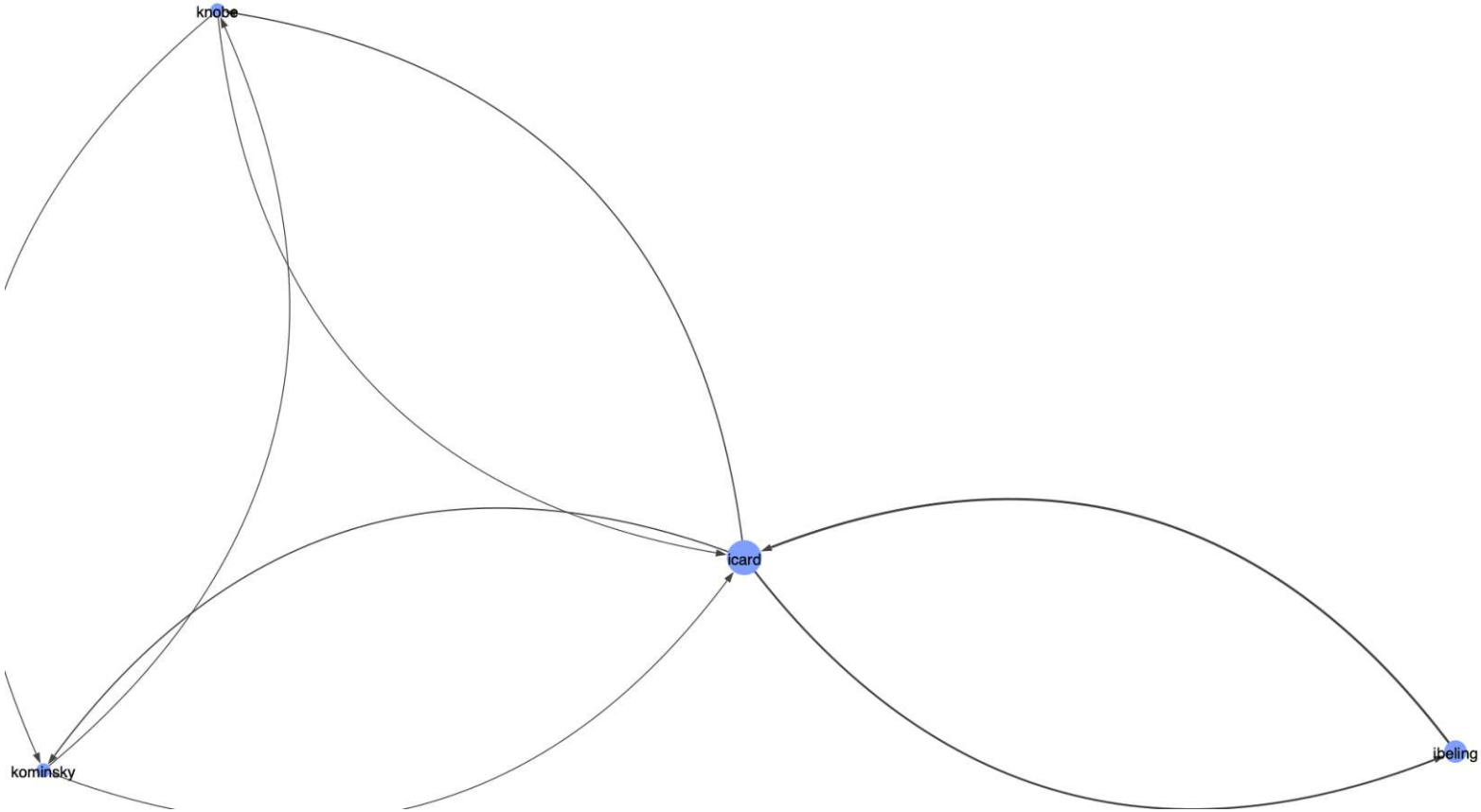


Figure 6.9: A subgraph of $\mathcal{G}_m$, depicting Group 8.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Figures

2.1 ER-Diagram of the database. 11

2.2 Number of publications and relevant publications added during each snowballing step. (From left to right $\mathcal{S}_x^c/\mathcal{S}_x$: 8/37, 79/204, 165/486, 7/30, 25/63, 3/7, 7/45) 18

2.3 Number of relevant publications per year (a negligible amount publications occur before 1970) 19

2.4 $\mathcal{G}_f$ where dark red (dark blue) indicates a(n) (ir)relevant publication in $\mathcal{S}_0$ and light red (light blue) indicates a(n) (ir)relevant publication in $\mathcal{S} \setminus \mathcal{S}_0$. (Isolated vertices are not depicted in this graph and edge direction is suppressed.) . 21

2.5 The publication graph $\mathcal{G}_p$, containing all relevant publications published after (and including) 2010, which consists out of 107 vertices (8 are in $\mathcal{S}_0^r$; 29 are in $\mathcal{S}_{-1}^r$; 42 are in $\mathcal{S}_{-2}^r$; 7 are in $\mathcal{S}_{+1}^r$; 17 are in $\mathcal{S}_{+2}^r$; 3 are in $\mathcal{S}_{+1-1}^r$; 1 are in $\mathcal{S}_{+2-1}^r$) and 326 edges. The vertex size correlates with vertex degree. 1: Graded Causation and Defaults; 2: Actual Causation: A Stone Soup Essay; 3: Cause without Default; 4: Actual Causation and the Art of Modeling; 5: Actual Causality in a Logical Setting; 6: Counterfactuals; 7: A Partial Theory of Actual Causation; 8: From Programs to Causal Models; 9: A Modification of the Halpern-Pearl Definition of Causality; 10: Explaining Actual Causation in Terms of Possible Causal Processes; (Isolated vertices are not depicted in this graph and edge direction is suppressed.) 22

2.6 The author graph $\mathcal{G}_a$ based on $\mathcal{G}_p$, which consists of 130 vertices and 462 edges. Darker colors indicate a higher number of publications in $\mathcal{G}_p$. Vertex size correlates with the weighted vertex degree; edge width correlates with edge weight. 1: Halpern; 2: Lagnado; 3: Vennekens; 4: Gerstenberg; 5: Hitchcock; 6: Bex; 7: Beckers; 8: Verheij; 9: Bochman; 10: Icard (Isolated vertices are not depicted in this graph.) 23

2.7 The collaboration graph $\mathcal{G}_c$, based on $\mathcal{G}_p$, which consists of 130 vertices and 192 edges. Darker colors indicate a higher number of publications in $\mathcal{G}_p$. Vertex size correlates with the weighted vertex degree. Edge width correlates with edge weight. An edge with color red has weight greater than 1. (1: Lagnado; 2: Eberhardt; 3: Gerstenberg; 4: Vennekens; 5: Zhang; 6: Goodman; 7: Tenenbaum; 8: Fontana; 9: Lee; 10: Lifschitz.) 24

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

2.8 The merged graph $\mathcal{G}_m$ based on $\mathcal{G}_p$, which consists of 130 vertices and 755 edges. Darker colors indicate a higher number of publications in $\mathcal{G}_p$. Vertex size correlates with the weighted vertex degree; edge width correlates with edge weight. 1: Lagnado; 2: Halpern; 3: Eberhardt; 4: Gerstenberg; 5: Vennekens; 6: Zhang; 7: Bex; 8: Tenenbaum; 9: Chockler; 10: Hitchcock. (Isolated vertices are not depicted in this graph and edge direction is suppressed.) . 25
2.9 A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_p$. 26
2.10 A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_a$. 26
2.11 A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_c$. 27
2.12 A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_m$ 27
2.13 A subgraph of $\mathcal{G}_m$, where the colours indicate community affiliation. 34
2.14 The subgraph from $\mathcal{G}_p$ induced by S_A 35

6.1 A line graph depicting the average in-degree, out-degree and overall degree of the publications in $\mathcal{G}_p$. 143
6.2 A subgraph of $\mathcal{G}_m$, depicting Group 1. 144
6.3 A subgraph of $\mathcal{G}_m$, depicting Group 2. 145
6.4 A subgraph of $\mathcal{G}_m$, depicting Group 3. 146
6.5 A subgraph of $\mathcal{G}_m$, depicting Group 4. 147
6.6 A subgraph of $\mathcal{G}_m$, depicting Group 5. 148
6.7 A subgraph of $\mathcal{G}_m$, depicting Group 6. 149
6.8 A subgraph of $\mathcal{G}_m$, depicting Group 7. 150
6.9 A subgraph of $\mathcal{G}_m$, depicting Group 8. 151

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

List of Tables

2.1 General properties of the discussed graphs. Other common measures such as average path length, radius and diameter, as well as vertex- and edge connectivity are omitted as all graphs in question are disconnected. 20
2.2 Degree Statistic of $\mathcal{G}_p$, $\mathcal{G}_a$ and $\mathcal{G}_c$. 20
2.3 Communities Overview 29
2.4 Top 15 authors according to the Degree Centrality, the In- and Out-Degree Centrality, the Betweenness Centrality, the Page Rank algorithm and the number of publication. 31
2.5 Top 15 publications according to the Degree Centrality, the In- and Out-Degree Centrality, the Betweenness Centrality and the Page Rank algorithm. . . . 32

3.1 Depicts which publication discuss which languages families 41
3.2 Summary of languages found in $\mathcal{F}$. This table categorises the languages based on whether they allow for quantification over variables; whether the variables used are multi-valued; whether the language is equipped with some form of default reasoning or is capable of encoding statements about normality; whether the language makes time explicit or whether it particularly emphasises sequences of events; whether the causal relations (or variables) can be probabilistic. (A check mark means that it fully satisfies the property; A tilde means that it partially satisfies the property; A question mark means that it is unknown if property is satisfied) 42
3.3 Depicts which publication discuss which token causality definition (before 2011). 59
3.4 Depicts which publication discuss which token causality definition (from 2011 onwards). 60
3.5 This table depicts which definitions rely on which language family 61
3.6 Summary of the token causality definitions discussed. The approaches considered are the Counterfactual approach (CF); a Process Oriented approach (PO); the Regularity Theoretic approach (RE); a (explicit) Probabilistic approach (PR). (*: Original in 2001; °: Original in 2003) 62
3.7 This table summarises which paper in $\mathcal{F}$ discusses which example 65

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

4.1 This table summarises the results obtained by applying the selected formalisms to the Benchmarks in Section 3.3. Each cell contains the literals (or formulas) that are deemed causes by the respective definition. To enhance readability the check mark indicate that the particular definition complies with intuitively "correct" answer for the scenario captured in the benchmark. The cases where no cause is present are indicated using ∅. 136

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Algorithms

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

[AB09] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[ACHI17] Gadi Aleksandrowicz, Hana Chockler, Joseph Y Halpern, and Alexander Ivrii. The computational complexity of structure-based causality. *Journal of Artificial Intelligence Research*, 58:431–451, 2017.

[AR16] Stefano V Albrecht and Subramanian Ramamoorthy. Exploiting causality for selective belief filtering in dynamic bayesian networks. *Journal of Artificial Intelligence Research*, 55:1135–1178, 2016.

[Bau13] Michael Baumgartner. A regularity theoretic approach to actual causation. *Erkenntnis*, 78(1):85–109, 2013.

[BDR+19] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

[BET11] Gerhard Brewka, Thomas Eiter, and Mirosław Truszczyński. Answer set programming at a glance. *Communications of the ACM*, 54(12):92–103, 2011.

[BG00] Chitta Baral and Michael Gelfond. Reasoning agents in dynamic domains. In *Logic-based artificial intelligence*, pages 257–279. Springer, 2000.

[BHM09] Helen Beebee, Christopher Hitchcock, and Peter Menzies. *The Oxford handbook of causation*. Oxford University Press, 2009.

[BJO18] Christer Bäckström, Peter Jonsson, and Sebastian Ordyniak. Novel structural parameters for acyclic planning using tree embeddings. In *IJCAI*, pages 4653–4659, 2018.

[BJT19] Francis Bloch, Matthew O Jackson, and Pietro Tebaldi. Centrality measures in networks. *Available at SSRN 2749124*, 2019.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[BL01] Phillip Bonacich and Paulette Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social networks*, 23(3):191–201, 2001.

[BL15] Alexander Bochman and Vladimir Lifschitz. Pearl's causality in a logical setting. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[Boc03] Alexander Bochman. A logic for causal reasoning. In *IJCAI*, pages 141–146, 2003.

[Boc04] Alexander Bochman. A causal approach to nonmonotonic reasoning. *Artificial intelligence*, 160(1-2):105–143, 2004.

[Boc18a] Alexander Bochman. Actual causality in a logical setting. In *IJCAI*, pages 1730–1736, 2018.

[Boc18b] Alexander Bochman. On laws and counterfactuals in causal reasoning. In *Sixteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2018.

[Bri12] Rachael Briggs. Interventionist counterfactuals. *Philosophical studies*, 160(1):139–166, 2012.

[BS17] Thomas Blanchard and Jonathan Schaffer. Cause without default. *Making a difference*, pages 175–214, 2017.

[BS18] Vitaliy Batusov and Mikhail Soutchanski. Situation calculus semantics for actual causality. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[BV12] Sander Beckers and Joost Vennekens. Counterfactual dependency and actual causation in cp-logic and structural models: a comparison. In *STAIRS*, volume 241, pages 35–46, 2012.

[BV15] Sander Beckers and Joost Vennekens. Combining probabilistic, causal, and normative reasoning in cp-logic. In *2015 AAAI Spring Symposium Series*, 2015.

[BV16] Sander Beckers and Joost Vennekens. A general framework for defining and extending actual causation using cp-logic. *International Journal of Approximate Reasoning*, 77:105–126, 2016.

[BV18] Sander Beckers and Joost Vennekens. A principled approach to defining actual causation. *Synthese*, 195(2):835–862, 2018.

[BVKPV10] Floris J Bex, Peter J Van Koppen, Henry Prakken, and Bart Verheij. A hybrid formal theory of arguments, stories and criminal evidence. *Artificial Intelligence and Law*, 18(2):123–152, 2010.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[CF17] Anthony Constantinou and Norman Fenton. Towards smart-data: Improving predictive accuracy in long-term football team performance. *Knowledge-Based Systems*, 124:93–104, 2017.

[CF18] Yoichi Chikahara and Akinori Fujino. Causal inference in time series via supervised learning. In *IJCAI*, pages 2042–2048, 2018.

[CFKL15] Hana Chockler, Norman Fenton, Jeroen Keppens, and David A Lagnado. Causal analysis for attributing responsibility in legal cases. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 33–42, 2015.

[CGS+18] Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9, 2018.

[CH10] Tom Claassen and Tom Heskes. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems*, pages 415–423, 2010.

[Che76] Peter Pin-Shan Chen. The entity-relationship model—toward a unified view of data. *ACM transactions on database systems (TODS)*, 1(1):9–36, 1976.

[Che19] Daniel L Chen. Judicial analytics and the great transformation of american law. *Artificial Intelligence and Law*, 27(1):15–42, 2019.

[CN+06] Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695(5):1–9, 2006.

[CQZ+19] Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery with cascade nonlinear additive noise models. In *IJCAI*, 2019.

[CXMR07] Peng Chen, Huafeng Xie, Sergei Maslov, and Sidney Redner. Finding scientific gems with google's pagerank algorithm. *Journal of Informetrics*, 1(1):8–15, 2007.

[DBV18] Marc Denecker, Bart Bogaerts, and Joost Vennekens. Causal reasoning in a logic with possible causal process semantics. In *17th INTERNATIONAL WORKSHOP ON NON-MONOTONIC REASONING NMR 2018*, pages 90–98. AAAI Press 2018, 2018.

[DBV19] Marc Denecker, Bart Bogaerts, and Joost Vennekens. Explaining actual causation in terms of possible causal processes. In *European Conference on Logics in Artificial Intelligence*, pages 214–230. Springer, 2019.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[dPMGAO11] Mónica del Pozo, Conrado Manuel, Enrique González-Arangüena, and Guillermo Owen. Centrality in directed social networks. a game theoretic approach. *Social Networks*, 33(3):191–200, 2011.

[DYFC09] Ying Ding, Erjia Yan, Arthur Frazho, and James Caverlee. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243, 2009.

[EL02] Thomas Eiter and Thomas Lukasiewicz. Complexity results for structure-based causality. *Artificial Intelligence*, 142(1):53–89, 2002.

[EW10] Martin Erwig and Eric Walkingshaw. Causal reasoning with neuron diagrams. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 101–108. IEEE, 2010.

[FG17] Luke Fenton-Glynn. A proposed probabilistic extension of the halpern and pearl definition of 'actual cause'. *The British journal for the philosophy of science*, 68(4):1061–1124, 2017.

[GDG+10] Clark Glymour, David Danks, Bruce Glymour, Frederick Eberhardt, Joseph Ramsey, Richard Scheines, Peter Spirtes, Choh Man Teng, and Jiji Zhang. Actual causation: a stone soup essay. *Synthese*, 175(2):169–192, 2010.

[GL10] Tobias Gerstenberg and David A Lagnado. Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, 115(1):166–171, 2010.

[GLL+04] Enrico Giunchiglia, Joohyung Lee, Vladimir Lifschitz, Norman McCain, and Hudson Turner. Nonmonotonic causal theories. *Artificial Intelligence*, 153(1-2):49–104, 2004.

[Hal] Joseph Y Halpern. Actual causality: A survey: Joseph halpern.

[Hal04] Ned Hall. Two concepts of causation. *Causation and counterfactuals*, pages 225–276, 2004.

[Hal07] Ned Hall. Structural equations and causation. *Philosophical Studies*, 132(1):109–136, 2007.

[Hal08] Joseph Y Halpern. Defaults and normality in causal structures. In *KR*, pages 198–208, 2008.

[Hal15a] Joseph Halpern. A modification of the halpern-pearl definition of causality. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Hal15b] Joseph Y Halpern. Cause, responsibility and blame: a structural-model approach. *Law, probability and risk*, 14(2):91–118, 2015.

[Hal16a] Joseph Y Halpern. *Actual causality*. MiT Press, 2016.

[Hal16b] Joseph Y Halpern. Appropriate causal models and the stability of causation. *The Review of Symbolic Logic*, 9(1):76–102, 2016.

[Hau05] Daniel Murray Hausman. Causal relata: Tokens, types, or variables? *Erkenntnis*, 63(1):33–54, 2005.

[HBF+19] Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. *IJCAI19*, 2019.

[HH59] Herbert Lionel Adolphus Hart and Tony Honoré. *Causation in the Law*. OUP Oxford, 1959.

[HH11] Joseph Y Halpern and Christopher Hitchcock. Actual causation and the art of modeling. *arXiv preprint arXiv:1106.2652*, 2011.

[HH15] Joseph Y Halpern and Christopher Hitchcock. Graded causation and defaults. *The British Journal for the Philosophy of Science*, 66(2):413–457, 2015.

[HHEJ13] Antti Hyttinen, Patrik O Hoyer, Frederick Eberhardt, and Matti Jarvisalo. Discovering cyclic causal models with latent variables: A general sat-based procedure. *arXiv preprint arXiv:1309.6836*, 2013.

[Hid05] Eric Hiddleston. Causal powers. *The British journal for the philosophy of science*, 56(1):27–59, 2005.

[Hit01] Christopher Hitchcock. The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98(6):273–299, 2001.

[Hit07a] Christopher Hitchcock. Prevention, preemption, and the principle of sufficient reason. *The Philosophical Review*, 116(4):495–532, 2007.

[Hit07b] Christopher Hitchcock. What's wrong with neuron diagrams. *Causation and explanation*, 4:69, 2007.

[Hit09] Christopher Hitchcock. Structural equations and causation: six counterexamples. *Philosophical Studies*, 144(3):391–401, 2009.

[Hit11] Christopher Hitchcock. Trumping and contrastive causation. *Synthese*, 181(2):227–240, 2011.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[HK09] Christopher Hitchcock and Joshua Knobe. Cause and norm. *The Journal of Philosophy*, 106(11):587–612, 2009.

[HP01] Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach: Part i: Causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 194–202, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[HP03] Mark Hopkins and Judea Pearl. Clarifying the usage of structural models for commonsense causal reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, pages 83–89. AAAI Press Menlo Park, CA, 2003.

[HP05] Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 56(4):843–887, 2005.

[HSJ17] Antti Hyttinen, Paul Saikko, and Matti Järvisalo. A core-guided approach to learning optimal causal graphs. In *IJCAI*, pages 645–651, 2017.

[Hum48] David Hume. *An enquiry concerning human understanding*. Open Court Press, 1748.

[II18] Duligur Ibeling and Thomas Icard. On the conditional logic of simulation models. *arXiv preprint arXiv:1805.02859*, 2018.

[II20] Duligur Ibeling and Thomas Icard. Probabilistic reasoning across the causal hierarchy. *arXiv preprint arXiv:2001.02889*, 2020.

[IKK17] Thomas F Icard, Jonathan F Kominsky, and Joshua Knobe. Normality and actual causal strength. *Cognition*, 161:80–93, 2017.

[Jel07] KA Jellinger. Causal models: How people think about the world and its alterations. *European Journal of Neurology*, 14(2):e17–e17, 2007.

[JZB18] Amin Jaber, Jiji Zhang, and Elias Bareinboim. A graphical criterion for effect identification in equivalence classes of causal diagrams. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 5024–5030. AAAI Press, 2018.

[KF08] Joshua Knobe and Ben Fraser. Causal judgment and moral judgment: Two experiments. *Moral psychology*, 2:441–8, 2008.

[KLRS17] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[KOP19] Martin Kronegger, Sebastian Ordyniak, and Andreas Pfandler. Backdoors to planning. *Artificial Intelligence*, 269:49–75, 2019.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[KS20] Shakil M. Khan and Mikhail Soutchanski. Necessary and sufficient conditions for actual root causes. In Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang, editors, *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 800–808. IOS Press, 2020.

[Kup16] Andrey Kupriyanov. Causality-based verification. 2016.

[LBV19] Emily LeBlanc, Marcello Balduccini, and Joost Vennekens. Explaining actual causation via reasoning about actions and change. In *European Conference on Logics in Artificial Intelligence*, pages 231–246. Springer, 2019.

[Lew74] David Lewis. Causation. *The journal of philosophy*, 70(17):556–567, 1974.

[Lew86] David Lewis. *Philosophical papers II*. Oxford: Oxford University Press, 1986.

[Lew13] David Lewis. *Counterfactuals*. John Wiley & Sons, 2013.

[LFWZ19] Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang. Context-aware emotion cause analysis with multi-attention-based neural network. *Knowledge-Based Systems*, 174:205–218, 2019.

[LG17] David A Lagnado and Tobias Gerstenberg. Causation in legal and moral reasoning. *The Oxford Handbook of Causal Reasoning*, pages 565–601, 2017.

[LJE+17] Quan Liu, Hui Jiang, Andrew Evdokimov, Zhen-Hua Ling, Xiaodan Zhu, Si Wei, and Yu Hu. Cause-effect knowledge acquisition and neural association model for solving a set of winograd schema problems. In *IJCAI*, pages 2344–2350, 2017.

[LLLY10] Joohyung Lee, Yuliya Lierler, Vladimir Lifschitz, and Fangkai Yang. Representing synonymity in causal logic and in logic programming. 2010.

[LSW19a] Rūta Liepiņa, Giovanni Sartor, and Adam Wyner. Evaluation of causal arguments in law: the case of overdetermination. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 214–218, 2019.

[LSW19b] Ruta Liepina, Giovanni Sartor, and Adam Z. Wyner. Arguing about causes in law: a semi-formal framework for causal arguments. *Artificial Intelligence and Law*, pages 1–21, 2019.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[LSW20] Rūta Liepiņa, Giovanni Sartor, and Adam Wyner. Arguing about causes in law: a semi-formal framework for causal arguments. *Artificial intelligence and law*, 28(1):69–89, 2020.

[LWFZ18] Junli Lu, Lizhen Wang, Yuan Fang, and Jiasong Zhao. Mining strong symbiotic patterns hidden in spatial prevalent co-location patterns. *Knowledge-Based Systems*, 146:190–202, 2018.

[LWZ17] Ruxia Liang, Jianqiang Wang, and Hongyu Zhang. Evaluation of e-commerce websites: An integrated approach under a single-valued trapezoidal neutrosophic environment. *Knowledge-Based Systems*, 135:44–59, 2017.

[LY10] Vladimir Lifschitz and Fangkai Yang. Translating first-order causal theories into answer set programming. In *European Workshop on Logics in Artificial Intelligence*, pages 247–259. Springer, 2010.

[LYF18] Jonathan Laurent, Jean Yang, and Walter Fontana. Counterfactual resimulation for causal analysis of rule-based models. In *IJCAI*, pages 1882–1890, 2018.

[M+07] Peter Menzies et al. Causation in context. *Causation, physics, and the constitution of reality*, pages 191–220, 2007.

[Mac65] John L Mackie. Causes and conditions. *American philosophical quarterly*, 2(4):245–264, 1965.

[McD95] Michael McDermott. Redundant causation. *British Journal for the Philosophy of Science*, pages 523–544, 1995.

[MGZ08] Nan Ma, Jiancheng Guan, and Yi Zhao. Bringing pagerank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008.

[MH69] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, 1969. reprinted in McC90.

[Min07] Marvin Minsky. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster, 2007.

[MR08] Sergei Maslov and Sidney Redner. Promise and pitfalls of extending google's pagerank algorithm to citation networks. *Journal of Neuroscience*, 28(44):11103–11105, 2008.

[MT+97] Norman McCain, Hudson Turner, et al. Causal theories of action and change. In *AAAI/IAAI*, pages 460–465, 1997.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Mu18] Kedian Mu. Measuring inconsistency with constraints for propositional knowledge bases. *Artificial Intelligence*, 259:52–90, 2018.

[MVDT00a] David Makinson and Leendert Van Der Torre. Input/output logics. *Journal of philosophical logic*, 29(4):383–408, 2000.

[MvdT00b] David Makinson and Leendert W. N. van der Torre. Input/output logics. *J. Philos. Log.*, 29(4):383–408, 2000.

[NFLG19] Martin Neil, Norman Fenton, David Lagnado, and Richard David Gill. Modelling competing legal arguments using bayesian model comparison and averaging. *Artificial Intelligence and Law*, 27(4):403–430, 2019.

[NJFD14] Michal Nykl, Karel Ježek, Dalibor Fiala, and Martin Dostal. Pagerank variants in the evaluation of citation networks. *Journal of Informetrics*, 8(3):683–692, 2014.

[Pau93] Gabriele Paul. Approaches to abductive reasoning: an overview. *Artificial intelligence review*, 7(2):109–152, 1993.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[Pea95] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[Pea98] Judea Pearl. On the definition of actual cause. 1998.

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.

[PM18] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[PY82] Christos H Papadimitriou and Mihalis Yannakakis. The complexity of facets (and some facets of complexity). In *Proceedings of the fourteenth annual ACM symposium on Theory of computing*, pages 255–260, 1982.

[RB08] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[Rei80] Raymond Reiter. A logic for default reasoning. *Artificial intelligence*, 13(1-2):81–132, 1980.

[Rei01] Raymond Reiter. *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. MIT press, 2001.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[Rei13] Daniel Reisberg. *The Oxford handbook of cognitive psychology*. Oxford University Press, 2013.

[Run12] Thomas A Runkler. Data analytics. *Wiesbaden: Springer. doi*, 10:978–3, 2012.

[Sch16] Jonathan Schaffer. Grounding in the image of causation. *Philosophical studies*, 173(1):49–100, 2016.

[Sch19] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[SG19] Dhanya Sridhar and Lise Getoor. Estimating causal effects of tone in online debates. In *IJCAI*, 2019.

[SGSH00] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[Shu11] Katrin Shulz. "if you'd wiggled a, then b would've changed": Causality and counterfactual conditionals.". *Synthese*, 179:239–251, 2011.

[Sim55] Herbert A. Simon. Causality and econometrics: Comment. *Econometrica*, 23(2):193–195, 1955.

[Smi81] Linda C Smith. Citation analysis. 1981.

[SOM17] Adam Summerville, Joseph Osborn, and Michael Mateas. Charda: Causal hybrid automata recovery via dynamic analysis. *arXiv preprint arXiv:1707.03336*, 2017.

[SP18] Andrew Selbst and Julia Powles. "meaningful information" and the right to explanation. In *Conference on Fairness, Accountability and Transparency*, pages 48–48. PMLR, 2018.

[SPG18] Dhanya Sridhar, Jay Pujara, and Lise Getoor. Scalable probabilistic causal structure discovery. In *IJCAI*, pages 5112–5118, 2018.

[Spi12] Daniel Spielman. Spectral graph theory. In *Combinatorial scientific computing*, number 18. Citeseer, 2012.

[SR15] Santiago Segarra and Alejandro Ribeiro. Stability and continuity of centrality measures in weighted graphs. *IEEE Transactions on Signal Processing*, 64(3):543–555, 2015.

[SSS+19] Shiv Shankar, Daniel Sheldon, Tao Sun, John Pickering, and Thomas G Dietterich. Three-quarter sibling regression for denoising observational data. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5960–5966, 2019.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[TK11] Charles R Twardy and Kevin B Korb. Actual causation by probabilistic active paths. *Philosophy of Science*, 78(5):900–913, 2011.

[VBD10] Joost Vennekens, Maurice Bruynooghe, and Marc Denecker. Embracing events in causal modelling: Interventions and counterfactuals in cp-logic. In *European Workshop on Logics in Artificial Intelligence*, pages 313–325. Springer, 2010.

[VDB09] Joost Vennekens, Marc Denecker, and Maurice Bruynooghe. Cp-logic: A language of causal probabilistic events and its relation to logic programming. *Theory and practice of logic programming*, 9(3):245–308, 2009.

[vdZLT19] Benito van der Zander, Maciej Liśkiewicz, and Johannes Textor. Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. *Artificial Intelligence*, 270:1–40, 2019.

[Ven11] Joost Vennekens. Actual causation in cp-logic. *Theory and Practice of Logic Programming*, 11(4-5):647–662, 2011.

[Ver17] Bart Verheij. Proof with and without probabilities. *Artificial Intelligence and Law*, 25(1):127–154, 2017.

[Wes15] Brad Weslake. A partial theory of actual causation. 2015.

[Wet18] Linda Wetzel. Types and tokens. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018. Accessed: 2020-03-30.

[WG17] Richard W Wright and Richard Goldberg. The ness account of natural causation: A response to criticisms. *Critical Essays on 'Causation and Responsibility*, pages 13–66, 2017.

[WLC18] Wei Wenjuan, Feng Lu, and Liu Chunchen. Mixed causal structure discovery with application to prescriptive pricing. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5126–5134. AAAI Press, 2018.

[Woh14] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pages 1–10, 2014.

[Woo05] James Woodward. *Making things happen: A theory of causal explanation*. Oxford university press, 2005.

[Wri87] Richard W Wright. Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts. *Iowa L. Rev.*, 73:1001, 1987.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[XM19] Zhipeng Xie and Feiteng Mu. Boosting causal embeddings via potential verb-mediated causal patterns. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1921–1927. AAAI Press, 2019.

[XWY+19] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Achieving causal fairness through generative adversarial networks. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1452–1458. ijcai.org, 2019.

[ZB17] Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780, 2017.

[ZHZ+17] Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, page 1347. NIH Public Access, 2017.

[ZL15] Haodi Zhang and Fangzhen Lin. Characterizing causal action theories and their implementations in answer set programming: Action languages b, c, and beyond. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[ZL17] Haodi Zhang and Fangzhen Lin. Characterizing causal action theories and their implementations in answer set programming. *Artificial Intelligence*, 248:1–8, 2017.

[ZLW+18] Zan Zhang, Lin Liu, Hao Wang, Jiuyong Li, Daning Hu, Jiaqi Yan, Rene Algesheimer, and Markus Meierer. Collective behavior learning by differentiating personal preference from peer influence. *Knowledge-Based Systems*, 159:233–243, 2018.

[ZS15] Dangzhi Zhao and Andreas Strotmann. Analysis and visualization of citation networks. *Synthesis lectures on information concepts, retrieval, and services*, 7(1):1–207, 2015.

[ZWW16] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*, 2016.

[ZWW18] Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving non-discrimination in prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 3097–3103. AAAI Press, 2018.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[ZZE+19] Zhalama, Jiji Zhang, Frederick Eberhardt, Wolfgang Mayer, and Mark Junjie Li. Asp-based discovery of semi-markovian causal models under weaker assumptions. In *IJCAI*, 2019.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.