

CAUSATION AND MODALITY



Dean McHugh

Causation and Modality

Dean McHugh

ILLC Dissertation Series DS-2023-05



INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

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

Copyright © 2023 by Dean McHugh.

Cover design and photograph by Dean McHugh; Inis Díomáin, 5 July 2020.
Printed and bound by Ipskamp Printing.

ISBN: 978-94-6473-115-6

Causation and Modality
Models and Meanings

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op dinsdag 23 mei 2023, te 12.00 uur

door Dean Michael McHugh
geboren te Dublin

Promotiecommissie

<i>Promotor:</i>	prof. dr. S.J.L. Smets	Universiteit van Amsterdam
<i>Copromotor:</i>	dr. K. Schulz	Universiteit van Amsterdam
<i>Overige leden:</i>	prof. dr. H.O. Dijstelbloem	Universiteit van Amsterdam
	prof. dr. J.M. Mooij	Universiteit van Amsterdam
	prof. dr. A. Kratzer	University of Massachusetts Amherst
	dr. M.D. Aloni	Universiteit van Amsterdam
	dr. F. Roelofsen	Universiteit van Amsterdam
	dr. I.A. Ciardelli	Università degli Studi di Padova
	dr. T.F. Icard III	Stanford University

Faculteit der Geesteswetenschappen



The research for/publication of this doctoral thesis received financial assistance from the Dutch Research Council (NWO) under the PhDs in the Humanities research programme (grant number [PGW.18.028](#)).

Between foreseeing and averting change
Lies all the mastery of elements
— Adrienne Rich

*For my parents,
Vivienne and Michael*

Contents

Acknowledgments	xv
1 Introduction	1
1.1 Motivations	1
1.1.1 Motivation 1. To assess the truth of explanations	1
1.1.2 Motivation 2. Legal certainty	4
1.2 Our answer to the modelling question	10
1.2.1 Formal model construction	13
1.3 Our answer to the meaning question	14
2 On the meaning of <i>cause</i> and <i>because</i>	17
2.1 Introduction	17
2.1.1 Why choose to analyse <i>cause</i> and <i>because</i> ?	17
2.1.2 The relation of <i>cause</i> and <i>because</i>	18
2.2 Sufficiency	20
2.2.1 The status of sufficiency	23
2.3 Difference-making: counterfactual dependence	26
2.4 Production	27
2.4.1 Sufficiency for production	30
2.5 Difference-making: the general pattern	33
2.5.1 The Perfection Principle	35
2.5.2 The ubiquity of the Perfection Principle	39
2.5.3 On the pragmatic origins of the Perfection Principle	39
2.5.4 Adding the Perfection Principle	40
2.6 Cause = difference-making + sufficiency + production	43
2.7 The modal force of difference-making	45
2.7.1 The implicatures of difference-making	47
2.7.2 <i>Only because</i>	49
2.7.3 Why counterfactual dependence is so compelling	51

3	Imaginative structures	53
3.1	Introduction	53
3.2	Comparing conditionals and causal claims	54
3.2.1	Sufficiency and hypothetical reasoning	56
3.2.2	The modal force of <i>will/would</i> and <i>cause/because</i>	59
3.2.3	A way out: selection functions	61
3.2.4	Adding time	63
3.3	A change of world	65
3.3.1	Coarser representations of the image	69
3.3.2	Proposal: varying a state	71
3.3.3	Simplifying the definition of what stays the same	71
3.4	The varieties of parthood	77
3.4.1	Imagination in an atomless world	80
3.4.2	Dependence between properties	81
3.4.3	Imagination without variables	85
3.5	Case study: state spaces of colour	89
3.6	Hypothetical scenarios raised by a sentence	93
3.6.1	A single-state definition of sentence variants	94
3.6.2	Adding time	96
3.6.3	Imagining changes to the laws	100
3.6.4	The intervention time	102
3.7	Exploring the present proposal	107
3.7.1	What is aboutness?	108
3.7.2	Aboutness for random variables	113
3.7.3	Capturing sufficiency violations	114
3.8	The boundless imagination	116
3.9	Conclusion	120
4	Imagining logically complex sentences true	123
4.1	Conditional inference patterns	124
4.1.1	Invalidating antecedent strengthening	124
4.1.2	Rational monotonicity	125
4.1.3	Cautious monotonicity	128
4.1.4	Reciprocity	130
4.1.5	Attempting to invalidate cautious monotonicity on the ordering approach	131
4.2	Simplification	133
4.2.1	Against Universal Realisability of the Antecedent	138
4.2.2	Cases supporting simplification	140
4.2.3	First-order clauses	148
4.2.4	Putting formal conditions in the restrictor	151
4.2.5	Applying the first-order clauses	151
4.3	Comparison with similarity approaches and premise semantics	155

4.3.1	Kratzer's situation-based semantics of conditionals	158
4.4	Reconstructing the present proposal within the ordering approach	160
4.4.1	Sufficiency in Kratzer's semantics of conditionals	164
4.4.2	From lumping to overlap	166
4.4.3	Definitions of overlap	168
4.4.4	Proving the equivalence of the ordering semantics and the present approach	170
4.5	Conclusion	172
5	An analysis of production	173
5.1	Introduction	173
5.2	Motivating production	174
5.2.1	Beckers' analysis of production	176
5.2.2	Sufficiency in Beckers' account of production	177
5.3	Production in terms of sufficiency	178
5.3.1	What is a chain?	178
5.4	Proper and improper chains with <i>cause</i> and <i>because</i>	179
5.5	The chain's relata: situations	181
5.5.1	Counterfactual dependence	183
5.5.2	Linking the chain	183
5.5.3	Evidence for this choice of aboutness: backtracking	186
5.5.4	Evidence for strong dependence: chain widening	186
5.6	Inevitable effects	189
5.6.1	Previous responses to the problem of inevitable effects . .	191
5.6.2	For every overdetermination case there is a switch	192
5.6.3	<i>Only</i> + <i>because</i> = counterfactual dependence	195
5.6.4	<i>Only because</i> on the present account	197
5.6.5	<i>Only because</i> an evidence for sufficiency	203
6	Causal models	205
6.1	Where we are	205
6.1.1	A formal model construction	206
6.1.2	On duration in causal models	207
6.1.3	Illustrating the model construction	208
6.2	Partial models and model abstraction	209
6.2.1	Examples of abstractions	213
6.2.2	The interaction between abstraction and possibility	214
6.2.3	Partial models of nomic possibility	216
6.2.4	A global view	221
6.3	Representing causal asymmetry	223
6.3.1	The asymmetry of <i>cause</i>	224
6.3.2	Alleged cases of simultaneous causation	224
6.3.3	Kim's cases	226

6.4	The concept of nomic possibility	229
6.4.1	Comparison with proof-theoretic views of laws of nature	231
6.4.2	The generality of nomic possibility	234
6.4.3	Representing models of physics	235
6.4.4	Comparison with alternatives to nomic possibility	237
6.5	A dynamic interpretation of structural causal models	240
6.5.1	The dynamic interpretation	241
6.5.2	The need for the dynamic interpretation	245
6.5.3	Previous hints at the dynamic interpretation	249
6.5.4	The exogenous/endogenous distinction	250
6.5.5	Transfinite dynamic interpretations	251
6.5.6	Recursive models and eventual truth	252
6.6	Interventions as sufficiency claims	254
6.6.1	From sufficiency to <i>would</i> -conditionals	260
6.6.2	Comparing interventions and <i>would</i> -conditionals	261
6.6.3	Probabilistic dynamic interpretations	263
6.7	Dense causal chains	265
6.7.1	The reality of dense dependence	267
6.7.2	An example: a light system	268
6.7.3	Dense causal chains in the present framework	269
6.7.4	The impossibility of dense causal chains in structural causal models	272
6.7.5	The impossibility of dense causal chains in Bayesian networks	274
6.7.6	Diagnosing the difference between structural causal models and the present framework	275
7	Exhaustification in the semantics of <i>cause</i> and <i>because</i>	279
7.1	Introduction	280
7.1.1	Three properties of <i>cause</i> and <i>because</i>	280
7.1.2	The apparent dissimilarity of properties 1, 2 and 3	281
7.1.3	Preliminaries: overview of the semantics of modality	282
7.2	<i>Cause</i> , <i>because</i> , and exhaustification	282
7.2.1	The simplified semantics	283
7.2.2	Properties 1, 2, and 3 via exhaustification	284
7.2.3	The full semantics	285
7.2.4	Why put exhaustification in the semantics of <i>cause</i> and <i>because</i> ?	286
7.2.5	Comparing the full and simplified semantics	288
7.3	The positive and negative conditions have the same background	289
7.3.1	The circumstances as modal base	291
7.3.2	Testing whether the positive and negative backgrounds can differ	293
7.4	Economy	297

7.4.1	<i>Because</i> and economy: data	298
7.4.2	<i>Because</i> and economy: analysis	299
7.5	Conclusion	300
Samenvatting		327
Abstract		329

Acknowledgments

First and foremost, to my supervisor, Katrin Schulz.

Katrin, this thesis would not have been possible without your brilliant insight and practical support. I count myself lucky to have had your continued encouragement, stretching back to our very first project together in January 2017. Then I sent you a paper which contained my first toy causal model, of an AND-gate (which somehow managed to sneak into this thesis all these years later), but also contained a single(!) citation – Frank Veltman’s *Logic of Conditionals*, a good choice but perhaps not enough to carry a whole research paper – and too many *Luftschlösser*. Undeterred, from then to now you have always been generous with your time, advice, and keen ability to see right to the heart of a problem. The way you blend disparate approaches, bouldering over disciplinary walls with ease, makes you a true Renaissance researcher, a perfect fit to advise on the Renaissance subject that is causality. You are a trailblazer, a role model, an inspiration.

I would also like to thank my promotor, Sonja Smets, for adeptly managing administrative matters and liaising with the NWO. And to my committee – Maria Aloni, Ivano Ciardelli, Huub Dijkstra, Thomas Icard, Angelika Kratzer, Joris Mooij, and Floris Roelofsen – thank you for agreeing to take up this task. Your ideas have been a tremendous source of insight, a debt shown in abundance throughout this long and often winding work.

I wish to thank the ILLC community as a whole for fostering a stimulating and interdisciplinary atmosphere. To my officemates in Science Park 107, Room 1.11 – Marco Degano, Levin Hornisher, Arnold Kochari, Ivar Kolvoort, Nadine Theiler, and Kaibo Xi – for creating a welcoming office. I owe a special gratitude to Levin. When I joined the PhD programme I expressly requested joining whichever office you were in – a decision in hindsight more wise than I could have imagined. I have always relished our conversations on logic, dynamics, and the combinatorics of CWI salad. I am also grateful to the members of the ILLC office: Jenny Batson, Caitlin Boonstra, Roos Bouwdewijn, Karine Gigengack, Tanja Kassenaar, Debbie Klaassen, Peter van Ormondt, and Marco Vervoort.

For riveting conversations about logic and causation throughout the years, I am grateful to Sander Beckers, Franz Berto, Nick Bezhanishvili, Justin Bledin, Lucas Champollion, Alexandre Cremers, Bridget Copley, Alison Fernandes, Luca Incurvati, Mingya Liu, Adrian Ommundsen, Jeroen Smid, Naftali Weinberger, and Bonan Zhao. A special word of thanks goes to Lucas for organising an online reading group on linguistics and philosophy which greatly influenced the direction of my research. I wish to also thank Giovanni Cinà and Adam Izdebski for teaching me much about causal modelling in healthcare, and Nicolien Janssens, Nikki Weststeijn, and the unforgettable, unforgotten Thijmen Nuninga for discussing causation in the law.

For their friendship, I thank Bahareh Afshari, Bahram and Nasrin Assadian, Lwenn Bussière, Fausto Carcassi, Julian Chingoma, Samuel Debray, Nadica Denić, Émile Enguehard, Swapnil Ghosh, Marianna Girlando, Saúl Fernandez Gonzalez, Jan Gronwald, Jacqueline Harding, Kristoffer Kalavainen, Henry Kalter, Artemii Korolkov, Alina Leidinger, Jan Maly, Erin McCloskey, Matteo Micheli, Yvette Oortwijn, Robert Paßmann, Mina Young Pedersen, Daira Pinto Prieto, Thomas Randriamahazaka, Allie Richards, Tom Roberts, Niccolò Rossi, Thomas Schindler, Julian Schlöder, Patrick Weigert, and Bastiaan van der Weij. I thank Ronald de Haan for those winter morning swims in the IJ, Olena Nahorna and Stephanie Rich for the laughter, and Angelica Hill, Giorgio Sbardolini, Aybüke Özgün, and Sebastian Speitel for their help organising academic events.

I have been fortunate to live with some truly remarkable people during the Master of Logic and PhD: in Botterweg 14 (the ‘Logic House’), Grzegorz Lisowski, Morwenna Hoeks, Jonathan Pesetsky, Zhuoye Zhao, Sam Adam-Day, Robin Martinot, Rachael Colley, Leo Lobski, Maëlle Havelange, Simon Vonlanthen, Anna Dmitrieva, Maximilian Siemers, and Jason Tsiaxiras; and in Haverstede 19 (the ‘Logic House 2.0’), Robin Martinot, Swapnil Ghosh, Laura Bussi, and Jonathan Melger. Morwenna, Jonathan, and Zhuoye, I treasure our long conversations on life and linguistics. Thanks especially to Swapnil for help copy editing this thesis (all remaining errors are still mine). And to Anna, thank you for the fond memories time-travelling through the chess multiverse.

To the ERC Choir, a dose of tranquility, a sweet source of gladness in the centre of Amsterdam. I am grateful in particular to call Elisabeth IJmker, Emma Pedley, Jared Penner, Anja Polstra, Nina Schroeder, and Kate Wareham my friends.

A special word of thanks goes to Milica Denić. Thank you for being so generous with your friendship and support. I have always treasured your advice on life, semantics, and rollerskating.

To Lorenzo Pinton, *grasie*. Your friendship has been a profound source of comfort and joy these past years. Thanks also to Marina and Diego for your hospitality in the most serene, Amsterdam of the South.

To my paranympths, Robin Martinot and Simon Rey. I cannot think of two better people to join me at the defense, since so many of the ideas contained within

were first discussed with you. Thank you for being dear friends and confidants.

Finally, to my family. To my brothers, Ben, Aaron, Ross, Ali, Abdul, Nizar, and Imran for creating a lively house, and Zara for the hospitality. Thanks especially to Evia, Rosie, and Nicole for offering a welcome distraction while I finished this thesis. And last but not most of all, to my parents. Thank you for being a constant source of support, for teaching me curiosity and kindness, and in challenges, defiance. This is for you.

Diemen
April 2023.

Dean McHugh

This thesis aims to answer two questions in the analysis of causation.

The modelling question. What information do we use when we judge that a causal claim holds? In other words, what information should a causal model contain?

The meaning question. Under what conditions is a causal claim true or false? That is, what do causal claims mean?

In this thesis, the causal claims we focus on are sentences containing *cause* or *because*. To take a simple example, suppose Alice flicks a light switch. The light turns on, and consider:

- (1)
 - a. Alice flicking the switch caused the light to turn on.
 - b. The light turned on because Alice flicked the switch.

What must things be like for such sentences to be true? What do they really say?

1.1 Motivations

Beyond intellectual curiosity, I would like to mention two reasons why these questions are worth pursuing.

1.1.1 Motivation 1. To assess the truth of explanations

People and public institutions alike aspire to give reasons for their actions. Aristotle believed that this capacity to give reasons goes to the heart of what it means to be human, that to be a human is to be a rational animal. As rational animals, we seek explanations. This is a cornerstone of public life. The transition from monarchs and dictators to elected representatives and judges is in part the transition from those who do not need to explain their actions to those who do. The

need for explanation is all the more pressing today given the rise of artificial intelligence (AI), bringing with it a need for explainable AI. For example, the EU's data protection law has responded to the rise of AI by affording a right to "the existence of automated decision making and ... meaningful information about the logic involved".¹ And just as we have a right to, say, water and privacy, Kate Vredenburg (2022) has recently argued that we have a right to explanation.

Now, it is all very well and good to have Enlightenment ideals, to demand reason and explanation. But it is a hollow demand without an ability to separate true explanations from false explanations. Without it, institutions would be free to broadcast whatever explanations they please, without reality getting in the way.

We typically give reasons and explanations using causal claims, using, for example, the words *cause* and *because*. It is well-known that these words have a rich and complex meaning. So determining whether an explanation is true or false is often a tricky matter. To illustrate, imagine a bank that decides to give out loans based on savings and years of higher education.² To get a loan of 10,000, an applicant needs 2,000 in savings and at least three years of higher education. Someone applies for a loan with three years of higher education but no savings. They are denied the loan. They go down to the bank to find out why.

CUSTOMER: Why was I denied the loan?

BANK MANAGER: Your application was denied because you have no savings.

CUSTOMER: Are you saying that if I had 2,000 in savings, I would have gotten the loan?

BANK MANAGER: That's right.

CUSTOMER: I had some savings, but I spent all the money on college. If I hadn't gone to college, sure, I would have 2,000 in savings, but then I wouldn't have three years of higher education and I still wouldn't have gotten the loan. So was I really denied the loan because I have no savings?

BANK MANAGER: Yes, as I said, your application was denied because you have no savings.

CUSTOMER: But I just told you that if I had kept my savings, I still wouldn't have gotten the loan. Telling me that I should have kept my savings is not useful advice.

BANK MANAGER: ???

¹GDPR Section 2, Article 13.2(f). For discussion see Wachter, Mittelstadt, and Russell (2017) and Kaminski (2019).

²I am grateful to Levin Hornischer for discussions of this example.

The conversation appears to break down, leaving the customer unsatisfied with their explanation. They might even file a complaint that their newfound right to explanation has been violated.

The customer's reasoning is an example of *backtracking*: when they imagine having 2,000 in savings, they take into account what the world would have to be like for that to be true. The bank manager, in contrast, uses non-backtracking reasoning: when they imagine the customer with 2,000 in savings they imagine the education level the same. They ignore the relationship between the customer's education level and the amount they have saved. In Chapter 3 we present a framework that can represent both kinds of reasoning, tracing their disagreement to different ways of resolving an ambiguity in how we construct hypothetical scenarios (specifically, an ambiguity in *when* to imagine the world changed so that the customer has 2,000 in savings; see section 3.6.4).

Here is a second example to illustrate the importance of being able to assess the truth of explanations. In 2015 the Dutch pension fund for government and education employees (the APB; i.e. my pension fund, and perhaps yours too) wrote a report on their approach to climate change. The report concludes with the following paragraph.

Would it not be better for ABP to get out of fossil fuels altogether?

Some people believe that selling all of our fossil fuel investments would send a powerful signal that we were taking climate change seriously. In practice, however, the effect would be minimal. Fossil energy companies would remain financially attractive investments and other investors would simply take our place. We might give a powerful signal but we would lose our influence at the same time. We believe we can do more on the climate change front by continuing to invest in these companies and encouraging them to become more sustainable.

Figure 1.1: From the [ABP's 2015 report on climate change](#).

“Other investors would simply take our place.” Why would the APB report mention this? One way to express the thought that the APB wish to implicitly communicate here is:

- (2) Our investment in fossil fuel companies is not causing increased fossil fuel emissions.

It is commonly argued that causation is required for moral responsibility (e.g. Braham and van Hees 2012, Beckers 2021c, though see Sartorio 2004 for an alternative view). If so, then the truth of (2) would absolve the ABP of moral responsibility for the increase in fossil fuel emissions that result from their investment. To assess the ABP’s culpability, then, we should determine whether (2) is true or false.

In the literature on causation, this is an example of the well-known problem of *overdetermination*; noted, for example, by Elisabeth Anscombe in her inaugural lecture:

It is not quite clear what ‘dependence’ is supposed to be, but at least it seems to imply that you would not get the effect without the cause. The trouble about this is that you might from some other cause. That this effect was produced by this cause does not at all show that it could not, or would not, have been produced by something else in the absence of this cause.

(Anscombe 1971:24)

In Chapter 2.4 we give an analysis of the meaning of *cause* and *because* where *C cause E* and *E because C* can be true, even though had the cause not occurred, the effect would have occurred anyway. The upshot is that we can pinpoint where exactly the argument from “it would have happened anyway without me” to “I didn’t cause it” breaks down. Moreover, we can do so using general observations about ordinary, apolitical scenarios – observations we can agree on regardless of our attitude toward climate change.

1.1.2 Motivation 2. Legal certainty

Given the centrality of causal notions to everyday life, it is unsurprising that many laws contain the words *cause* and *because*. While Hart and Honoré (1959) distinguish ‘causation in fact’ and ‘legal causation’, many judges abide by the plain meaning rule, which states that if ordinary meaning of a statute is clear, the judge must interpret it in that way.³ In courts that abide by the plain meaning rule, the present work on the ordinary meaning of *cause* and *because* can help resolve legal disputes about those words.

One legal area abundant in causal language is discrimination law. As an example, take Title VII of the US Civil Rights Act of 1964, which provides:

It shall be an unlawful employment practice for an employer to fail or refuse to hire or to discharge any individual ... because of such individual’s race, color, religion, sex, or national origin.

³The plain meaning rule also goes by the name textualism. Famously, Elena Kagan, a Justice of the US Supreme Court, [quipped in 2015](#) that “we’re all textualists now” (referring to US jurisprudence).

Section 703(a)(1), p. 255

The meaning of this statute hinges on the meaning of *because*. In ordinary life it is often clear what *because* means, but cracks in our understanding appear under the weight of legal scrutiny. Take the 2020 US Supreme Court case, *Bostock v. Clayton County*. Gerald Bostock worked for Clayton County, Georgia. In 2013 he joined a gay softball league and mentioned it at work. A few weeks later he was fired for “conduct unbecoming a county employee” ([Court opinion](#), p. 3). Bostock took his employer to court, arguing that the firing was illegal under Title VII.

At the time, Georgia – like most other US states – had no state law protecting against employment discrimination on the basis of sexual orientation (see Figure 1.2). However, the Civil Rights Act of 1964 was federal law, and therefore applied in Georgia. It mentioned sex discrimination, but, as one might expect from a law written in 1964, makes no mention of sexual orientation.

At the time there was considerable disagreement as to whether discrimination on the basis of sexual orientation violates Title VII. The Second and Seventh Circuits decided that it does, while the Eleventh Circuit, which includes Georgia, decided that it does not.⁴ Hence the need for the Supreme Court to settle the question.

The question before the Supreme Court was whether Gerald Bostock was fired because of his sex, and more generally, whether firing someone because of their sexual orientation constitute firing because of their sex. One might initially think that, since sex and sexual orientation are distinct traits, it is possible to fire someone because of one trait without firing them because of the other.

Pamela Karlan, arguing for Gerald Bostock, disagreed. At oral argument she asked the court to imagine two employees who both mention that they married their respective partners, who are male, on the weekend. The boss gives the first employee some time off to celebrate, and fires the other ([Oral argument](#) pp. 7–8). Why the different treatment? The first was a woman who married a man, the second a man who married a man. And, Karlan argues, if two people do the same thing, with the only difference between them being that one is a woman and one

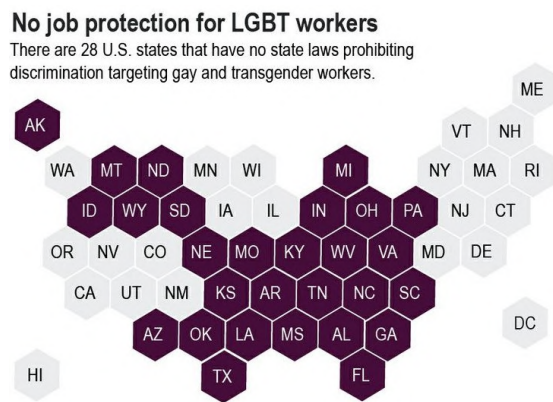


Figure 1.2: [The Associated Press](#), October 15, 2019

⁴The Second Circuit ruling comes from *Zarda v. Altitude Express, Inc.* (2018), the Seventh Circuit ruling from *Hively v. Ivy Tech Community College of Indiana* (2017), and the Eleventh Circuit ruling from *Evans v. Georgia Regional Hospital* (2017).

is a man – that is sex discrimination.



Figure 1.3: The cover of the *New York Times* the day *Bostock* was decided.

In a 6–3 majority, the Supreme Court agreed that discrimination on the basis of sexual orientation constitutes discrimination on the basis of sex, and therefore violates Title VII. The landmark ruling immediately extended employment protection to millions on LGBTQ Americans.

As soon as the Court’s Opinion was published, however, it came under attack. Some legal scholars have gone so far as to declare that “*Bostock* was bogus” (Berman and Krishnamurthi 2021). Brett Kavanaugh, in his dissent (with Clarence Thomas concur-

ring), accused the court of acting as an unelected legislature: “Instead of a hard-earned victory won through the democratic process, today’s victory is brought about by judicial dictate ... it was Congress’s role, not this Court’s, to amend Title VII.” Some legal scholars have argued that the Court’s interpretation in *Bostock* represents a dramatic shift in meaning.⁵

Did the majority amend Title VII or discover something it already entails? If we are to settle whether discrimination because of sexual orientation constitutes discrimination because of sex, we need to understand what *because* really means. The standard legal test for causation is the *but-for* test; in legal parlance, *but for the cause, the effect would not have occurred*. In contemporary terms: if the cause had not occurred, the effect would not have occurred.⁶ As Justice Elena Kagan put it at oral argument:

KAGAN: What you do when you look to see whether there is [sex]

⁵Some examples:

Bostock fundamentally redefined what it means to discriminate because of sex, expanding the definition of include discrimination based on an characteristic that is definitionally related to, and thus logically inseparable from, sex.

(Cohen 2022:407).

Bostock articulated a new mixed motive theory that allows a Title VII plaintiff to prove “but for” causation in cases where the employer acted partly for an impermissible reason and partly for a permissible reason so long as the impermissible reason was decisive.

(Cain 2021:464)

⁶For a discussion of the but-for test and its shortcomings in tort law (the law of injuries and accidents), see the Harvard Law Review (2017).

discrimination under Title VII is, you say, would the same thing have happened to you if you were of a different sex?

([Oral argument](#), pp. 41–42)

Under the but-for test, then, we have to ask whether sentence (3) is true.

(3) If Gerald Bostock had been a woman, he wouldn't have been fired.

When we imagine what would have happened if he were a woman, intuitively there are many possibilities to consider. (As Alito exclaimed during oral argument: “the parties have in their briefs, have all of these comparisons, and they will make your head spin if you – if you try to figure them all out!”) If Gerald Bostock were a woman, he could have been a woman who is attracted to men, in which case he would have kept his job, or he could have been a woman who is attracted to women, in which case he still would have been fired (given that the employer had a blanket rule against gay people in general). Samuel Alito picks up on this point in his dissent:

the Court carefully includes in its example just two employees, a homosexual man and a heterosexual woman, but suppose we add two more individuals, a woman who is attracted to women and a man who is attracted to women. . . . We now have the four exemplars listed below, with the discharged employees crossed out:

~~Man attracted to men~~
 Woman attracted to men
~~Woman attracted to women~~
 Man attracted to women

The discharged employees have one thing in common. It is not biological sex, attraction to men, or attraction to women. It is attraction to members of their own sex—in a word, sexual orientation. And that, we can infer, is the employer's real motive.

The Court tries to prove that “it is impossible to discriminate against a person for being homosexual or transgender without discriminating against that individual based on sex,” . . . but as has been shown, it is entirely possible for an employer to do just that. . . . discrimination because of sexual orientation or transgender status does not inherently or necessarily constitute discrimination because of sex.

([Justice Alito](#), pp. 16–17)

Kavanaugh's dissent makes the same point:

Consider the employer who has four employees but must fire two of them for financial reasons. Suppose the four employees are a straight

man, a straight woman, a gay man, and a lesbian. The employer with animosity against women (animosity based on sex) will fire the two women. The employer with animosity against gays (animosity based on sexual orientation) will fire the gay man and the lesbian. ... To treat one as a form of the other ... misapprehends common language, human psychology, and real life.

(Justice Kavanaugh, pp. 12–13)

Alito and Kavanaugh’s dissents emphasise the possibility that, had Bostock been a woman, he could have been attracted to women, in which case he still would have been fired. They use the existence of this possibility to argue that sex discrimination does not constitute sexual orientation discrimination. And sure enough, given this possibility, we cannot say that (3) is true. If Gerald Bostock had been a woman, he might have been fired, he might not. The but-for test appears to fail, or is at least inconclusive.

As we will see in this thesis, the but-for test is a poor approximation of the meaning of *because*. One reason for this, discussed in McHugh (2020), is that a *because* sentence can be true even when the cause is stronger than strictly required for the claim to hold. Take the following naturally-occurring examples.⁷

- (4) a. He has an American passport because he was born in Boston. [Source]
- b. Reyna received a Danish passport because her mother was born in Copenhagen. [Source: *The Bolton News*]

These sentences are perfectly acceptable. This judgement is something we can all agree on, regardless of our politics, judicial philosophy or attitude toward discrimination law.

Now look what happens when we apply the but-for test:

- (5) a. If he hadn’t been born in Boston, he wouldn’t have received an American passport.
- b. If Reyna’s mother hadn’t been born in Copenhagen, Reyna wouldn’t have received a Danish passport.

These are clearly unacceptable. When we imagine, say, Reyna’s mother not being born in Copenhagen, there are intuitively many places where she could have been born instead. In some of these cases, Reyna would still have received a Danish passport, in others not.

The conclusion I draw from these data (in McHugh 2023 and chapter 2 here) is that *because* does not require that, had the cause not occurred, in **every** case

⁷For further examples see McHugh (2020:§2). In section 2.7.1 we show that (4) can trigger a false inference (called an *implicature*) with emphasis on *Boston* and *Copenhagen*. This may lead one to mistakenly think they are false. Nonetheless, these sentences are true according to the meaning of *because* we propose.

we consider, the effect would not have occurred; rather, it is enough that if the cause had not occurred, in **some** case we consider the effect would not have occurred.⁸ In logical terminology, we may say that *because* has an existential, rather than universal difference-making condition. Accordingly, we may distinguish a ‘universal’ but-for test (the current legal standard) from an ‘existential’ but-for test.

The plain meaning rule – which Alito and Kavanaugh accept – requires interpreting Title VII according to the meaning of the words it actually contains; in this case, the meaning of *because*. The data in (4) show that our interpretation of Title VII will be more faithful to the meaning of *because* if we adopt the existential but-for test in place of the universal but-for test.

This switch has serious ramifications for the outcome of *Bostock*. All parties agree that if Gerald Bostock had been a woman, there is a possibility in which he would have kept his job; namely, if he had been a woman attracted to men. This is the possibility that Pamela Karlan’s argument and the Court opinion appeal to. This is not enough to pass the universal but-for test – as Alito and Kavanaugh’s dissents make abundantly clear, emphasising the possibility of Gerlard Bostock still being fired if he were a woman who is attracted to women. But it is enough to satisfy the existential but-for test, which as we have seen better reflects the meaning of *because*. According to the existential but-for test, *Bostock* was not “bogus”, but correctly decided.

This kind of argument is not unique to *Bostock*, nor indeed to sexual orientation discrimination. One case decided alongside *Bostock* concerned Aimee Stephens, a trans woman was fired after informing her employer that she wished to work in female clothing. The logic of the case is similar to *Bostock*. If Aimee Stephens were a different sex (which the court interpreted as: if she had been assigned a different sex at birth), intuitively she could have been cisgender, and she could have been transgender. In the former case she wouldn’t have been fired, which constitutes proof of sex discrimination according to the existential but-for test.

Similar concerns arise in pregnancy discrimination. Many countries have laws preventing sex discrimination but make no explicit provisions against pregnancy discrimination. To illustrate, take the landmark 1990 case of *Dekker v. VJV Centrum* from the European Court of Justice. The VJV Centrum was a youth centre in Amsterdam. Their insurer had a rule stating that they can refuse to pay for an employee’s absence if the employee is “unable to work within six months of the date on which the insurance commenced, when such inability to work within half a year was clearly to be anticipated from the state of health of the person concerned at the time when he commenced work” (see the [Court judgement](#)). Elisabeth Dekker applied for a job at the centre when she was three-

⁸Indeed, even this is more than a *because* sentence requires, as we discuss in section 2.4. Though there are some other aspects of the meaning of *because* we discuss in section 2.2.

months' pregnant. The employer acknowledged that she was the most qualified candidate, but refused to hire her given the risk that their insurer would refuse to pay for her replacement.

The employer argued in court that they did not refuse to hire her because she was pregnant, but because their insurer would refuse to pay for a replacement.⁹ It is not sex discrimination, they argued, since they would also refuse to hire any man in the same situation; that is, any man who needs to take an absence and whose replacement will not be paid by the insurer. This view was taken not only by the employer regarding this case but also by the UK government.¹⁰ One could imagine an employer making a similar argument in any pregnancy discrimination lawsuit, to argue against giving employees special provisions due to pregnancy.

This illustrates a general issue that has vexed discrimination law, known as the *comparator problem*.¹¹ If Elisabeth Dekker were a man, to whom should we compare her? A man who is fit to work? A man with a foreseeable absence?

If Elisabeth Dekker had been a man, she would likely have been fit to work. But we cannot rule out definitively that she would not have had a foreseeable absence, in which case the insurer would not have paid for a replacement. The existence of such a possibility proves sex discrimination on the existential but-for test but not on the universal but-for test. In a pregnancy discrimination case it may usually be granted that if the pregnant woman were a man, among the possibilities we consider is one where she would have been treated differently (for example, where she would have been hired). Given this, the existential but-for test, but not the universal but-for test, tells us that that pregnancy discrimination is a form of sex discrimination.

1.2 Our answer to the modelling question

Our answer to the modelling question is that a causal model must represent three things: time, part-whole relations, and nomic possibility. Following the tradition of possible-worlds semantics, we begin with the set of logically possible worlds. By 'representing time, part-whole relations, and nomic possibility', we mean the following.

⁹“the VJV did not offer the post to Mrs Decker, not because she was a woman nor because she was pregnant, but because the VJV’s insurer had informed it that in cases of leave on grounds of pregnancy or any subsequent unfitness for work which might be linked to pregnancy and confinement the Risicofonds [the insurer] could refuse to pay any benefits.” ([Court judgement](#), p. 3949).

¹⁰“The *United Kingdom* takes the view that ... a woman shall not be rejected for a post on the ground that she is or will become unable to work, when a man would not have been rejected on the same ground.” ([Court opinion](#), pp. 3949–50)

¹¹For an overview of the problems created by the need to find comparators, see Goldberg (2011).

Part-whole relations. For the model to represent part-whole relations (also known as mereological structure), we mean that the model contains a *state space*. A state space is a partial order where each element, called a *state* represents the state of a part of the world at a moment in time, and the order represents parthood. For example, the state of Amsterdam is part of the state of the Netherlands, the state of the tulip's colour is part of the state of the tulip, and the state of the logic institute is part of the state of the university. Each maximal element of the state space represents the state of the whole world at a moment in time.

To illustrate, Figure 6.4 represents the mereological structure of the switch and the light. State s is part of a state t just in case there is an upward line from s to t . For example, the state of the light being on is part of the state of the light being on and the switch down.

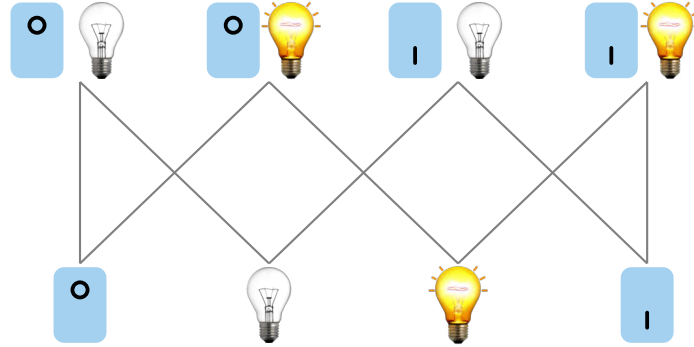


Figure 1.4: A state space of the switch and light.

Our model will use part-whole relations to capture how we construct hypothetical scenarios. In chapter 3 we propose that when we construct a hypothetical scenarios, we ‘remove’ a part of the world at intervention time, and consider all the worlds that contain what is left over. To determine this part of the world, and define what it means to remove one part of the world from another, we need to represent part-whole relations.

In section 3.4.3 we show that state spaces are strictly more general than a representation based on random variables, in the sense that every set of random variables can be represented by a state space, but not every state space can be represented by random variables.

Time. For the model to represent time, we mean that the model represents each possible world as a linear order, with each point of the order representing the state of the world at a moment in time, and the order representing time.¹² For example, Figure 1.2 depicts a world containing the switch and the light.

¹²For another representation of time where linear orders play a central role, see Maudlin (2014).



Figure 1.5: A nomically possible world.

Our model contains time since interpreting *cause* or *because* involves considering hypothetical scenarios, and on the present proposal, how we construct these hypothetical scenarios exhibits a temporal asymmetry: we pick a moment in time at which to imagine a change (what we call *intervention time*) and keep the past fixed but allow the future to vary. To capture this temporal asymmetry we need to represent time.

Nomic possibility. For the model to represent nomic possibility, we mean that the model determines for each possible world whether it is nomically possible or nomically impossible. For example, there is a nomically possible world in which the apple falls to the ground, but no nomically possible world where it spontaneously turns to gold. Formally, where W is the set of logically possible worlds, our model picks out a subset P of W , where a world is nomically if it is in P and nomically impossible if it is not.

For example, the world in Figure 1.5 is nomically possible. It represents the possibility of the switch being flicked, the light turning on, and then the switch being flicked up again. In contrast, Figure 1.6 depicts a world where the light spontaneously flickers on and off. This world is nomically impossible.



Figure 1.6: A nomically impossible world.

In addition to these three components, our model also contains two language-specific components: an aboutness relation and an interpretation function.

Aboutness. Our model contains a relation between sentences and states, telling us when a sentence is *about* a state. For example, “The switch is up” is about the state of the switch, and not about the state of the light.

In Chapter 3 we use the aboutness relation to represent what parts of the world we fix and what parts we vary when we interpret a causal claim. The idea is that to evaluate C *cause* E or E *because* C , we allow the part of the world C is about at intervention time to vary.

Interpretation function. Finally, our model includes an interpretation function, which for each sentence of our language returns the set of possible worlds in

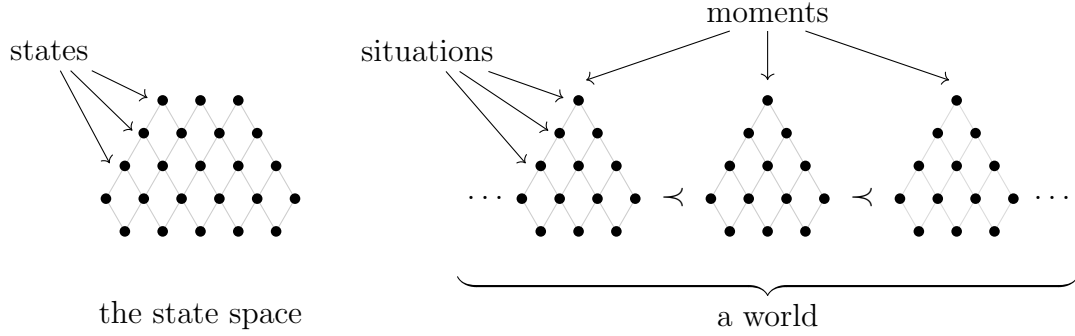


Figure 1.7: The relationship between states, situations, moments and worlds.

which it is true.

1.2.1 Formal model construction

Formally, we take a *state space* (S, \leq) to be a partially ordered set, the elements of which we call *states*, with \leq representing parthood. We assume that every state is part of a maximal state with respect to parthood.¹³ We take states to represent snapshots, representing how some things stand at a moment in time. We can construct worlds from states as follows.

- A *situation* is a particular instance of a state.
- A *moment* is a situation that is not part of any other situation.¹⁴
- A *world* is a linear order of moments.

We need situations since the same state may appear multiple times in a world. Formally, the set of situations, moments and worlds are defined, respectively, as follows.

$$\begin{aligned} Sit &:= S \times I, \text{ where } I \text{ is an arbitrary label set,} \\ M &:= \{t_i \in Sit : t \leq u \text{ implies } t = u \text{ for all } u \in S\}, \\ W &:= \{(M', \preceq) : M' \subseteq M, \preceq \text{ is a linear order}\}. \end{aligned}$$

Figure 1.7 illustrates this model construction.

Putting everything together, the models we use to interpret causal claims are defined as follows.

Given a set of sentences \mathcal{L} , a model is a tuple

$$(S, \leq, \mathcal{A}, P, |\cdot|)$$

¹³That is, we assume $\forall s \in S \exists t \in S : s \leq t \wedge \forall u \in S (t \leq u \Rightarrow t = u)$. In Fine's (2017) terminology, our state spaces are *world-spaces*.

¹⁴We define that situations inherit parthood relations from their states; that is, situation s_i is part of situation t_j just in case state s is part of state t .

where (S, \leq) is a state space, $\mathcal{A} \subseteq \mathcal{L} \times S$, $P \subseteq W$, and $|\cdot| : \mathcal{L} \rightarrow W$.

The intuitive interpretation of these components is that $\mathcal{A}(A, s)$ represents that sentence A is about state s , $w \in P$ that world w is nomically possible, and $w \in |A|$ that A is true at w .

1.3 Our answer to the meaning question

Our answer to the meaning question analyses the meaning of *cause* and *because* in terms of three relations: *sufficiency*, *production*, and *difference-making*.

Sufficiency. In Chapter 3 we analyse the notion of *sufficiency*, stating when the truth of one sentence is sufficient for the truth of another. We use sufficiency to capture the difference in meaning between, for example, the (a) and (b)-sentences below.

- (6) a. Ali being born in Europe caused him to get Irish citizenship.
b. Ali being born in Ireland caused him to get Irish citizenship.
- (7) a. Ali has an Irish passport because he was born in Europe.
b. Ali has an Irish passport because he was born in Ireland.

On our analysis, sufficiency inhabits a zone between logical entailment and material implication. Sufficiency is weaker than logical entailment. For example, in a context where there is electricity in the building, “Alice flicked the switch” is sufficient for “the light is on”, even though the former does not entail the latter since there is a logically possible world (say, a world where the power is out) where Alice flicks the switch and the light does not turn on. And sufficiency is stronger than material implication; for example, one can activate a random number generator and get a seven, even though activating the random number generator was not sufficient to get a seven. Similarly, Ali can be in fact born in Europe and have an Irish passport, without being born in Europe being sufficient to get an Irish passport.

Production. Following Hall (2004) and Beckers (2016), our semantics also includes a notion of production. Loosely, the truth of sentence C produces the truth of sentence E just in case there is a chain of counterfactual dependence from C to E . Chapter 5 is devoted to the analysis of production. We use production to account for the well-known overdetermination cases – cases where a causal claim is true even though, had the cause not occurred, the effect would have occurred anyway.

Difference-making. Finally, our semantics incorporates a notion of difference-making, as analysed by Sartorio (2005). Section 2.5 shows how to add Sartorio's notion of difference-making to any given semantics of *cause* and *because*.

The resulting semantics we propose is that *C cause E* and *E because C* are true just in case *C* is true, and *C* is sufficient for *C* to produce *E* but $\neg C$ is not. Where \gg represents sufficiency and *produce* production, we propose that *C cause E* and *E because C* are true just in case the following holds.

$$C \wedge (C \gg (C \text{ produce } E)) \wedge \neg(\neg C \gg (\neg C \text{ produce } E))$$

Chapter 2

On the meaning of *cause* and *because*

2.1 Introduction

In this chapter we analyse the meanings of English words *cause* and *because*.¹ We will work within the research programme of truth-conditional semantics (Davidson 1967b, Lewis 1970a): our task is therefore to determine under what conditions sentences containing *cause* or *because*, such as those in (1), are true or false.

- (1) a. Alice flicking the switch caused the light to turn on.
 b. The light turned on because Alice flicked the switch.

2.1.1 Why choose to analyse *cause* and *because*?

There are many other words in English with a causal flavour we could analyse instead, such as *therefore*, *since*, *for*, *due to*, *as*, *so*, *make*, *have*, *force* and *let*. We choose to analyse the verb *cause* since most work on the meaning of causal claims analyses this word, and we would like to compare our analysis with this work. And we choose to analyse *because* since it is used far more frequently than *cause*,² and does not carry some of the restrictions on use faced by *cause*.³ To

¹Parts of this chapter have been previously published in McHugh (2020), (2022) and (2023).

²Searches of the British National Corpus (BNC) and Corpus of Contemporary American English (CCAE) reveal that for every occurrence of either *a cause* or *the cause* there are approximately 3 occurrences of *caused* (in both the BNC and CCAE) and 36 (BNC) and 62 (CCAE) occurrences, respectively, of *because*. Frequency of *a cause*: 609 (BNC), 4852 (CCAE); *the cause*: 2161 (BNC), 16586 (CCAE); *caused*: 9243 (BNC), 62527 (CCAE); *because*: 99496 (BNC), 1346051 (CCAE). Corpora accessed at <https://www.english-corpora.org/bnc/> and <https://www.english-corpora.org/coca/> on 5 October 2020.

³Childers (2016:§3.3) argues on the basis of corpus and experimental evidence that there are two senses of *cause*, differing in register and sentiment. One is used in formal contexts—such as academic research—and can be used with any sentiment toward the effect, the other is informal but expresses a negative feeling toward the effect. Since *because* appears more frequently and with less constraints on its use than *cause*, our analysis will be more applicable if we analyse

illustrate, Figure 2.1 depicts the frequency of some causal words from Google Books’ English corpora from 1500 to 2019.⁴ The chart shows *because* firmly in the lead, *caused* and *causes* a distant second, *cause of* (which includes *a/the cause of* and *an actual cause of*, analysed by e.g. Pearl 2000, Hall 2004, Halpern 2016 and Beckers 2016) far behind, and *causally depends* (analysed by e.g. Lewis 1973a) in comparison hardly mentioned at all.

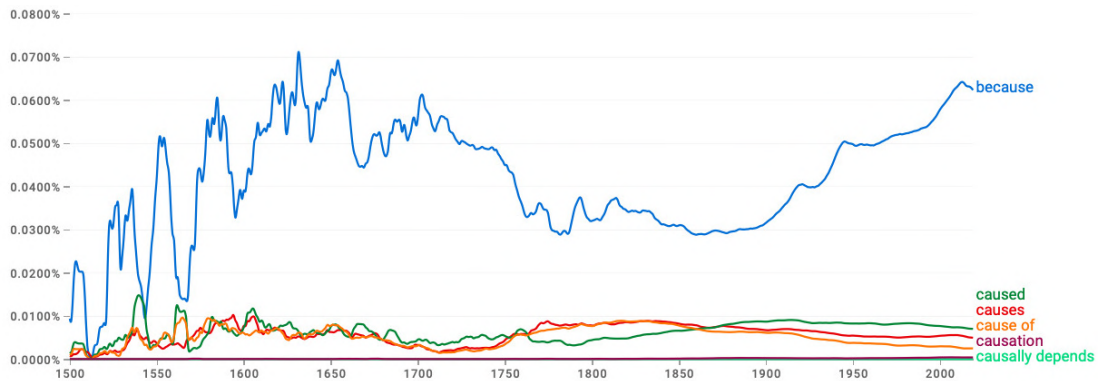


Figure 2.1: Frequency of *because*, *caused*, *causes*, *cause of*, *causation* and *causally depends* from the Google Books Ngram viewer [Source].

Our analysis will therefore enjoy greater applicability by focusing on words that are frequently used, such as *because* and *cause*.

2.1.2 The relation of *cause* and *because*

Let us briefly address the issue of the relation of *cause* and *because*. In the literature on causation one often finds lengthy discussions on the relation of causation (e.g. Ehring 1987, 2009, Hausman 2005, Moore 2005, Schaffer 2016). Here we are not analysing causation but causal claims; specifically, the meaning of *cause* and *because*. *Cause* is a verb whose subject is a noun phrase and whose object is a determiner phrase or a *to*-infinitive, and *because* is a preposition taking two clauses.⁵

because as well as *cause*.

⁴Though note that Google’s Ngram viewer is not always a reliable source for corpus work. For discussion see Younes and Reips (2019).

⁵Note that *because* is not conjunction, as *say*, *and* is. Some evidence for this is that one can prepose *because* but not conjunctions:

- (i) a. Because I ate too much pasta, I am sleepy.
- b. *And I ate too much pasta, I am sleepy.

Traditionally, *because* has been called a “subordinating conjunction” and *because*-clauses have been called “adverbial clauses” as have *after*, *although*, *before*, conditional *if*, *since*, *though*, *unless* and *while*. Jespersen (1924:89) proposed that they are in fact prepositions, a claim supported by Geis (1970), Huddleston and Pullum (2004), Pullum (2009, 2014).

While determining the relata of causation may raise special difficulties, determining the relata of causal claims does not. This is a consequence of a foundational assumption in semantics, the principle of compositionality: “The meaning of an expression is a function of the meanings of its parts and of the way they are syntactically combined” (Partee 1984; for an overview see Pagin and Westerstål 2010a,b). The principle of compositionality implies that expressions have the same meaning when embedded under *cause* or *because* as they have in other environments. There is therefore no special problem of relata of causal claims. The arguments of *cause* and *because* mean whatever they mean in general.

Comparing the semantics of *cause* and *because* requires assuming some relationship between their relata. For simplicity we will give a semantics of *cause* in terms of clauses, and assume that nouns and *to*-infinitives in some sense express clauses. For example, in (1a) the noun *Alice flicking the switch* expresses the clause *Alice flicked the switch* and the *to*-infinitive *the light to turn on* expresses the clause *the light turned on*.

This is a simplification since clauses differ from noun phrases and *to*-infinitives in a number of ways. As Vendler (1967) discusses, different kinds of nominalisation exhibit different relationships with the verb they contain. This gives rise to truth-conditional differences when paired with *cause*. To illustrate, imagine Gonzalo is playing roulette and bet on green. There is a $1/37 \approx 2.7\%$ chance of the ball landing on green. Against all the odds, it does. Consider (2).

- (2) a. The fact that the ball rolled on the wheel caused Gonzalo to win.
- b. The ball rolling on the wheel caused Gonzalo to win.
- c. The ball’s roll on the wheel caused Gonzalo to win.

To my ear, (2a) is unacceptable, (2b) is slightly better, and (2c) is fine. In the terminology of Vendler (1967:131), *the ball rolling on the wheel* is an ‘imperfect’ nominal, “in which the verb is still alive as a verb,” while *the ball’s roll on the wheel* is a ‘perfect’ nominal, “in which the verb is dead as a verb, having become a noun.” These differences can likely account for the contrasts in (2), say, by proposing that ‘the fact that’ and imperfect nominals denote event types (sets of events), while perfect nominals denote a single event.

One point worth clarifying is that the verb *cause* does not express a relation between events. One nit-picky reason for this is that, assuming the traditional distinction between states and events, *cause* can also relate states, as in (3).

- (3) The rose’s colour caused the little prince to be happy.

Linguists often use ‘eventuality’ as an umbrella term covering states and events alike. So one might think that *cause* expresses a relation between eventualities. What would it mean for *cause* to express a relation between eventualities? Presumably, it would mean we can associate each noun phrase and *to*-infinitive with an eventuality in a given context (the eventuality it in some sense denotes in a

given context), and there is a relation R between eventualities such that for any context w , c *cause* e is true in w just in case the eventuality denoted by c in w stands in relation R to the eventuality denoted by e in w .

Some evidence against this is that two noun phrases can intuitively denote the same eventuality but give rise to differences in meaning in a *cause* sentence. Consider:

- (4) a. Ali being born in Ireland caused him to receive an Irish passport.
- b. Ali being born in Europe caused him to receive an Irish passport.

There is a clear contrast in meaning between them. Arguably, however, given that Ali was in fact born in Ireland, the event of him being born in Ireland is identical to the event of him being born in Europe. If *cause* expressed a relation between eventualities, this contrast would be unexpected.

To make sure the pattern is robust, here is a second example. Imagine Fatima and Freddy are playing a game where they have to raise their hand after exactly one minute has passed. Freddy raises his hand after 50 seconds and loses; Fatima raises her hand after one minute and wins. Compare:

- (5) a. Fatima raising her hand caused her to win.
- b. Fatima raising her hand after one minute caused her to win.

In each case there is a difference in meaning between the (a)- and (b)-sentences. The (b)-sentences are much more acceptable than the (a)-sentences. Arguably, however, the event of Fatima raising her hand is the same event as her raising her hand after one minute.

2.2 Sufficiency

With these preliminaries out of the way, let's discuss the meaning of *cause* and *because* in earnest. Recall the sentences in (1), repeated below.

- (1) a. Alice flicking the switch caused the light to turn on.
- b. The light turned on because Alice flicked the switch.

Suppose Alice walks into a dark room and flicks a light switch. The light connected to the switch turns on. In this case we are perfectly happy to say the sentences in (1). If we wonder why they are true, a plausible response is that, if Alice hadn't flicked the switch, the light wouldn't have turned on. In other words, the light turning on counterfactually depended on Alice flicking the switch (where E counterfactually depends on C just in case, if C had not occurred, E would not have occurred).

The idea that the interpretation of causal claims involves counterfactual dependence is widespread, adopted by the panoply of so-called counterfactual de-

pendence approaches to causation (for an overview see Collins, Hall, and Paul 2004). The idea goes back to David Hume, who in 1748 wrote that “We may define a cause to be an object followed by another ... where, if the first object had not been, the second never had existed” (Hume 1748:§VII). Taking up this idea, Lewis (1973a) proposed that an event c causally depends on an event e just in case there is a chain of counterfactual dependence from c occurring to e occurring.

It is well-known that there are cases where a causal claim is true but the effect does not counterfactually depend on the cause (this is the problem of overdetermination, which we discuss in section 2.4). Another problem is that counterfactual dependence is not enough for the corresponding causal claim to be true.⁶ For example, imagine a robot has to get to Main Street, choosing between any of the four available routes to get there (see Figure 2.2).

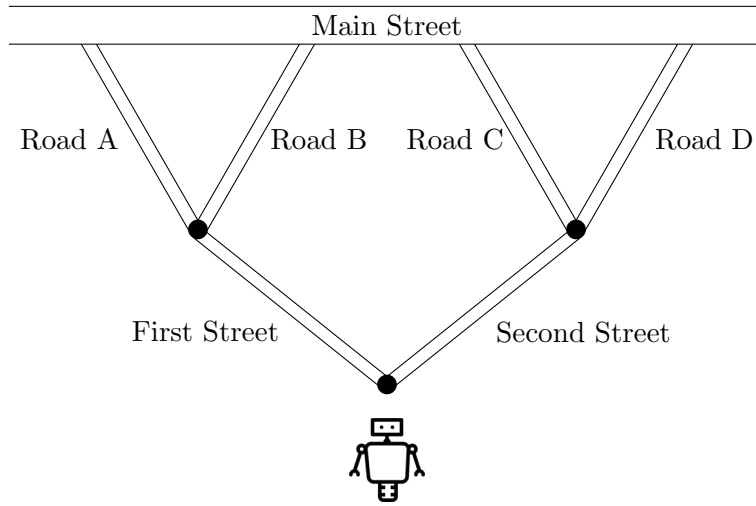


Figure 2.2

When the robot reaches a fork in the road, it decides what way to go at random. It is programmed so that it must take one of the routes to Main Street and cannot reverse. On this particular day, the robot took First Street and then Road B. Now consider the sentences below.

- (6) a. The robot taking First Street caused it to take Road B.
- b. The robot took Road B because it took First Street.

In this context, (6a) and (6b) are intuitively false. However, the robot taking Road B counterfactually depended on it taking First Street: if it hadn't taken

⁶That being said, Nadathur and Lauer (2020) propose that counterfactual dependence is not enough for a sentence containing *make* – such as “Gurung made the children dance” to be true. They do, however, propose that counterfactual dependence, together with the cause and effect both occurring, is enough for a sentence containing *cause* to be true.

First Street, it wouldn't have taken Road B. In general, then, counterfactual dependence is not enough for the truth of the corresponding causal claim.

Intuitively, the sentences in (6) are false because the robot did not have to take Road B after taking 1st Street. It could have taken Road A instead. This suggests that *cause* and *because* imply that the cause is in some sense sufficient for its effect (in other words, causes 'guarantee', 'determine', or 'ensure' that their effects happen).⁷

We can test this explanation by changing the context in a minimal way to make taking First Street sufficient for taking Road B, and checking whether our intuitions change accordingly. To that end, suppose that before setting out on its journey the robot was programmed with the rule: *Always change direction!*. The effect of the rule is that if the robot goes left at one fork in the road, it must go right at the next one, and if it goes right at the first fork, it must go left at the second. Today, the robot took First Street and, since it went left first, its programming required it to turn right, taking Road B. In this new context, consider again the sentences in (6).

Suddenly the sentences in (6) are true. Counterfactual dependence does not account for the contrast between the turn-at-random and always-change-direction contexts. For in both, if the robot hadn't taken First Street it wouldn't have taken Road B. The difference is rather a difference in sufficiency.

What exactly does it mean for the truth of one sentence to be sufficient for the truth of another? Anscombe remarks that "“sufficient condition” is a term of art whose users may ... lay down its meaning as they please" (Anscombe 1971:5). Chapter 3 gives an analysis of the notion sufficiency implied by *cause* and *because*. Our analysis predicts that when the robot turns at random, taking First Street is not sufficient to take Road B, but when it is programmed to always change direction, it is. Without going into too much detail here, these predictions result from incorporating the openness of the future into the analysis of sufficiency. Informally, let us say that for sentence *A* to be sufficient for sentence *C*, *C* must be true in every nomically possible future after *A* becomes true.⁸ When the robot turns at random, but not when it is programmed to always change direction, it is nomically possible for it to take Road A after First Street. If we then assume

⁷The claim that causes must be sufficient for their effects also helps makes sense of some statements from the history of philosophy. For example, Spinoza writes

(i) Nothing exists of which it cannot be asked, what is the cause (or reason) [*causa (sive ratio)*], why it exists. (*Principles of Cartesian Philosophy*, Part I, Axiom 11)

This is commonly taken as a statement of the principle of sufficient reason (see Melamed and Lin 2021). But if *cause* did not require sufficiency, it would be hard to see how this would count as a statement of the principle.

⁸Semantics of conditionals that use the openness of the future include Thomason (1970), Thomason and Gupta (1980), Arregui (2005), Ippolito (2013), Khoo (2015), and Canavotto (2020).

that *C cause E* and *E because C* imply that *C* is sufficient for *E*, we correctly predict (6) to be unacceptable when the robot turns at random but acceptable when it is programmed to always change direction.

Now, some violations of sufficiency do not involve the openness of the future. Consider (7)–(9).

- (7) *Ali was born in Ireland and has Irish citizenship.*
 - a. Ali got Irish citizenship because he was born in Europe.
 - b. The fact that Ali was born in Europe caused him to get Irish citizenship.
- (8) *Nina is 16 and tried to get into a bar for over 18s. The bouncer did not let her in.*
 - a. The bouncer refused to let Nina in because she is under 30.
 - b. The fact that Nina is under 30 caused the bouncer to refuse to let her in.
- (9) *Yves only buys a specific shade of blue paint: ultramarine. In a paint shop he sees paints in various colours, including various shades of blue. He bought some ultramarine paint.*
 - a. Yves bought this paint because it is blue.
 - b. The fact that this paint is blue caused Yves to buy it.

These sentences are intuitively unacceptable. They also significantly improve when we replace the cause with a minimally different one that is sufficient for the effect; e.g. replacing *Europe* with *Ireland*, *under 30* with *under 18*, and *blue* with *ultramarine*. This is exactly what we expect if *C cause E* and *E because C* imply that *C* is sufficient for *E*.

As in the robot case, these contrasts are not due to a difference in counterfactual dependence. For example, if Alice had been over 18, or over 30, she would have gotten in. Nor are they due to the openness of the future. Accounting for, say, (7a)'s unacceptability in this way would require a time when the paint was blue but not yet determined that it would be ultramarine. But specific shades do not come after general shades in time – every blue is simultaneously a specific kind of blue.

2.2.1 The status of sufficiency

So we have evidence that *cause* and *because* imply that the cause is sufficient for the effect. What is the nature of this inference? Is it an entailment, part of the literal meaning of *cause* and *because*? Is it a presupposition, something that we take for granted when we interpret these words? Or is it not part of the literal meaning, but inferred via pragmatic reasoning? In this section we provide evidence that the sufficiency inference is an entailment, rather than a

presupposition or implicature.

One of the simplest ways to test the status of an inference is to put it in a downward-entailing environment, such as under negation. To that end, consider:

- (10) a. The robot turned at random. Taking First Street didn't cause it to take Road B.
 b. The robot turned at random. It didn't take Road B because it took First Street.

These are intuitively acceptable. Indeed, it seems we can say something stronger:

- (11) a. Given that the robot turned at random, nothing caused it to take Road B.
 b. Given that the robot turned at random, it did not take Road B because of anything.

These are also intuitively fine. Now, given that the robot turned at random, nothing was sufficient for it to take Road B (this is something our analysis of sufficiency in Chapter 3 will predict). And if *C cause E* and *C because E* entail that *C* is sufficient for *E*, by contraposition, *C* not being sufficient for *E* entails $\neg(C \text{ cause } E)$ and $\neg(C \text{ because } E)$. Thus assuming that the sufficiency inference is an entailment, we correctly predict (10) and (11) to be true.

These data also show that the sufficiency inference is not a presupposition. It is standardly assumed that presuppositions project through negation: if *A* presupposes *B*, then $\neg A$ does too. For example,

- (12) Iris's brother doesn't have a car. \leadsto Iris has a brother.

Thus if *A* presupposes *B*, we expect $\neg B \wedge \neg A$ to be incoherent, since $\neg A$ implies *B*, contradicting $\neg B$.

- (13) # Iris doesn't have a brother. Her brother doesn't drive a car.

(13) is clearly incoherent. Now, if *C cause E* and *C because E* presupposed that *C* is sufficient for *E*, we would expect (10) and (11) to be similarly incoherent, since then $\neg(C \text{ cause } E)$ and $\neg(E \text{ because } C)$ would imply that *C* is sufficient for *E*, but the fact the robot turned at random implies that *C* was not sufficient for *E*. The stark contrast between the incoherence of (13) and acceptability of (10) and (11) is evidence that the sufficiency inference is not a presupposition.

Finally, (10) and (11) show that the sufficiency inference is not an implicature. It is standardly assumed that implicatures are not derived when doing so would lead to an overall weaker meaning (Chierchia 2013:129, Fox and Spector 2018). For example, the disjunction in *A or B* is typically strengthened to an exclusive

reading via scalar implicature, but under negation it is interpreted with its literal, inclusive meaning, as illustrated in (14) (where \leadsto denotes ‘intuitively implies’).

- (14) a. Fionn passed syntax or semantics.
 \leadsto Fionn passed syntax or semantics and not both.
 b. Fionn didn’t pass syntax or semantics.
 $\not\leadsto \neg(\text{Fionn passed syntax or semantics and not both}).$

If the sufficiency inference were an implicature, we would similarly expect it to disappear in downward entailing environments, such as under negation. As we saw in (10) and (11), when the cause is not sufficient for the effect, *cause* and *because* under negation are fine. Given that the unnegated sentences were unacceptable due to a failure of sufficiency, if the sufficiency inference were an implicature this would be unexpected.

Let’s see how *cause* and *because* under negation behave in a context where the cause is sufficient for the effect. Consider:

- (15) The robot was programmed to always change direction.
 a. Taking First Street didn’t cause it to take Road B.
 b. It didn’t take Road B because it took First Street.
 (16) Given that the robot was programmed to always change direction,
 a. nothing caused it to take Road B.
 b. It didn’t take Road B because of anything.

These sentences sound much worse in this context than when it turns at random. As discussed, the only salient difference between the two contexts appears to be a difference in sufficiency: when the robot taking First Street was sufficient for it to take Road B, the sentences are acceptable; when it is not, they are unacceptable. If the sufficiency inference were an implicature, we would expect it to disappear under negation. Then the difference in judgement between the two contexts would also disappear. But this is not what we observe.

In summary, data from *cause* and *because* under negation show that the sufficiency inference is an entailment, and not a presupposition or implicature.⁹

Before moving on, let us address one lingering feeling one may have about the status of sufficiency. One might think that the sufficiency inference is not part of the literal meaning of *cause* and *because*, arguing along the following lines: (i) the sufficiency violations in (6)–(9) sound unacceptable because they do not offer not a good explanation of why their effects occurred; (ii) what it means for an explanation to be good depends in part on pragmatic considerations; (iii) a good

⁹Section 5.6.5 provides further evidence that the sufficiency inference is not an implicature, based on the meaning of *only because*. Our argument there is a bit more complex, requiring an overview of the full semantics of *because* we propose, so for now we simply refer the reader to that section for further details.

explanation of why the effect occurred should account for why the effect occurred rather than not.

This is a compelling idea. Indeed, the present account wholeheartedly agrees with each of the points (i)–(iii) above. However, they do not imply the apparent conclusion that the sufficiency inference is not part of the literal meaning of *cause* and *because*. It is hard to say what it means in general for an explanation to be good: that depends greatly on context – whom one is talking to, what information can be taken for granted, the purpose of the conversation, and myriad other factors. Nonetheless, one requirement we can all agree on is that for an explanation to be good, it should at least be true. For example, if one says that the robot took Road B because it took First Street, for that to count as a good explanation the sentence should be true. So even with all the unclarity about what it means in general for an explanation to be good, by putting sufficiency into the literal meaning of *cause* and *because* we account for point (i): the feeling that (6)–(9) are not good explanations of why their effects. They are not good explanations because they are not true. Point (ii) is also correct, but the ‘in part’ is crucial: what it means to be an explanation to be depends on both semantic and pragmatic considerations. It just so happens in this case that we can account for point (i) by appealing to semantic considerations alone. Finally, point (iii) is correct since a *cause* or *because* claim in which the cause is not sufficient for the effect is not true, and therefore does not count as a good explanation.

2.3 Difference-making: counterfactual dependence

Of course, the semantics of *cause* and *because* involves more than just sufficiency. The cause must also in some sense make a difference to the effect. As Lewis put it, “We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it” (Lewis 1973a:557). This difference-making component is illustrated in the following scenario, due to Hall (2000) and depicted in Figure 2.3.

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

(Hall 2000:205)

Consider (17) in this context.

- (17) a. The train reached the station because the engineer flipped the switch.
- b. The engineer flipping the switch caused the train to reach the station.

The sentences in (17) are intuitively unacceptable.

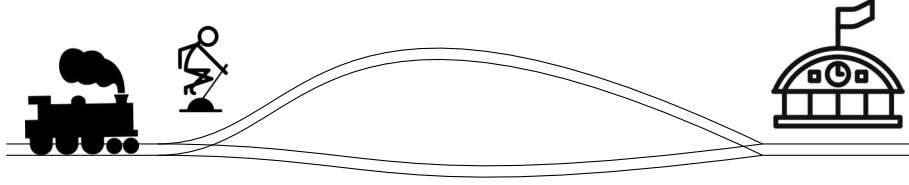


Figure 2.3: Hall's switching scenario.

Given the set up of the scenario, the train would have reached the station whether or not the engineer had flipped the switch. Accordingly, our analysis of sufficiency predicts that the engineer flipping the switch is sufficient for the train to reach the station, and the engineer not flipping the switch is also sufficient for the train to reach the station. Sufficiency alone is not enough for a *cause* or *because* claim to be true. It only considers what happens in cases where the cause is true. To account for the unacceptability of (17), it seems we must say something about what happens in cases where the cause is false. Let us call this the ‘difference-making requirement’ of *cause* and *because*.

What does the difference-making requirement consist in? A compelling thought is that (17) are unacceptable because even if the engineer hadn't flipped the switch, train would have reached the station anyway. As is well-known, the idea that causation requires counterfactual dependence is plagued by a host of counterexamples (see Lewis 2000, Hall and Paul 2003, Hall 2004, Halpern 2016, Beckers 2016, Andreas and Günther 2020 and many more). Let us turn to those counterexamples now, and then return for a fuller discussion of difference-making in section 2.5.

2.4 Production

Here is a much-discussed example introduced by (Hall and Paul 2003:110) (the following formulation is from Hall 2004:235).

Suzy and Billy, expert rock-throwers, are engaged in a competition to see who can shatter a target bottle first. They both pick up rocks and throw them at the bottle, but Suzy throws hers before Billy. Consequently Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's had not occurred, so the shattering is overdetermined.

Consider (18) and (19) in this context.

- (18) a. The bottle broke because Suzy threw her rock at it.

- b. Suzy throwing her rock at the bottle caused it to break.
- (19) a. The bottle broke because Billy threw his rock at it.
- b. Billy throwing his rock at the bottle caused it to break.

Intuitively, the sentences in (18) are acceptable but the sentences in (19) are not.

The difference between Suzy and Billy is not a difference in sufficiency. Suzy throwing her rock is sufficient for the bottle to break, and Billy throwing his rock is sufficient for the bottle to break (given that they are both expert rock-throwers). So if not sufficiency, where could the difference in judgement between (18) and (19) come from?

Foreseeing this concern, Anscombe (1971:24) writes,

It is not quite clear what ‘dependence’ is supposed to be, but at least it seems to imply that you would not get the effect without the cause. The trouble about this is that you might from some other cause. That this effect was produced by this cause does not at all show that it could not, or would not, have been produced by something else in the absence of this cause.

Notice how in discussing this problem, Anscombe appeals to a notion of production: “That this effect was *produced* by this cause does not at all show that it could not, or would not, have been *produced* by something else” (my emphasis). This suggests a way to distinguish Billy and Suzy in this case: Suzy throwing her rock produced the bottle to break, Billy throwing his rock did not. This strategy has been further discussed by Lewis (1986), Hall (2004), and Beckers (2016), among others. So we propose that *C cause E* and *E because C* entail that *C produced E*.

If this strategy is going to work, the big question is to say what production consists in. We take production to be a relation between sentences. The truth of one sentence produces the truth of another. What does it mean for sentence *A* to produce sentence *C*? Chapter 5 gives an analysis of production. The guiding idea is that *A* produces *C* just in case there is a chain of counterfactual dependence from the truth of *A* to the truth of *C*. Let us briefly introduce our analysis of production here, to show how we can make predictions about the Billy and Suzy case using it.

Our analysis is inspired by Lewis’s analysis of causal dependence (Lewis 1973a). The difference is that we use it to analyse a part of the meaning of causal claims – production – while Lewis used it to analyse the whole – causal dependence. (In other words, Lewis’s analysis is a case of mistaken synecdoche. Like a sergeant who, asked to put boots on the ground, brings a pile of footwear to battle, Lewis mistook the part for the whole.)

It is tempting to say that the chain is made up of *events*: *A* produces *C* just in case there is a chain of events, beginning with an *A*-event, ending with

a *C*-event, such that each event on the chain counterfactually depends on the previous event.

A major obstacle for an events-based analysis of production is the identity criteria of events are remarkably complex. We lack a general theory of what it means for events to be the same. To illustrate, Hitchcock (2012:83) wonders:

if a meeting is originally scheduled for Monday at noon, and then re-scheduled for Tuesday at noon, is the meeting that occurs on Tuesday at noon the very same meeting that would have occurred on Monday? That is, was the meeting postponed, strictly speaking, or was the original meeting cancelled and a different meeting scheduled for Tuesday?

This question is reminiscent of the ship of Theseus, and does not seem to admit a determinate answer. A further difficulty is that some have proposed a causal theory of the identity criteria of events. For example, Davidson (1969) proposes to individuate events by their causes and effects. We cannot appeal to such an account here on pain of circularity.

Given this predicament, an analysis of production that appealed to events would quickly get stuck with tricky questions about what events are, their identity criteria, and so on. This raises the question whether we must take the winding roads through events to get to production, or whether there is a bypass. I believe there is: a much clearer approach is to take the chain to consist of propositions holding at a particular time.

Recall that our model includes a state space, whose maximal states represent the state of a world at a moment in time. We will take a proposition to be a set of maximal states (intuitively, those at which the proposition is true).¹⁰ Briefly put, our analysis of production from Chapter 5 is that sentence *A* produces sentence *C* at a world *w* just in case *A* and *C* are true at *w*, and there is a chain of proposition–time pairs such that (i) for each element p_t of the chain, *p* is true at world *w* at *t* (ii) *A* implies the first proposition on the chain and *C* the last, (iii) the chain is temporally saturated, and (iv) each element of the chain counterfactually depends on a buffer of previous elements.¹¹

¹⁰Recall that we take a world to be a linear order of moments (the linear order representing time), a moment to be a maximal situation with respect to parthood, and a situation to be a particular instance of a state. We extend truth-at-a-state to truth-at-situation by saying that *p* is true at a moment t_i just in case *p* is true at the state *t*.

Note that by taking propositions to be sets of states, rather than sets of situations, the notion of proposition adopted here incorporates an *intrinsicness* requirement (discussed by Hall 2004) – whether a proposition is true at a moment depends only on the state of that moment, and not on the world in which the state happens to find itself. This means, for example, that the chain cannot contain an element like p_t where *p* expresses that Suzy threw the rock at some other time t' .

¹¹Formally, the chain is a set of proposition–time pairs that is linear, in the sense that for any elements p_t and $q_{t'}$ of the chain, $t \leq t'$ or $t' \leq t$. By the claim “counterfactually depends on a

On this analysis we can show that Suzy throwing her rock produced the bottle to break, but Billy throwing his rock did not. To show the former, consider a chain such as

Suzy throws her rock at t .
 Suzy's rock is flying toward the bottle at location x' at t' ...
 Suzy's rock is flying toward the bottle at location x'' at t'' .
 Suzy's rock hits the bottle at t''' .
 The bottle breaks at t'''' .

As the reader may verify, this counts as a production chain according to our definition, so we predict that Suzy throwing her rock produced the bottle to break.

Let's now show that Billy throwing did not produce the bottle to break. To my knowledge virtually all theories predicting that Billy's throw did not cause the bottle to break rely on the same observation; namely, that Billy's rock did not hit the bottle (though different accounts, e.g. Halpern 2016, Beckers 2016, incorporate this observation into their formal accounts in different ways). We can also use this to show that Billy did not produce the bottle to break on our analysis. For Billy's throw to produce the bottle to break, we would have to find a proposition p and time t such that there is a chain of counterfactual dependence running from Billy throwing when he did to p being true at t , and if p had not been true at t , the bottle would not have broken when it did. The only proposition with a chain of counterfactual dependence from Billy throwing to it would be something like "Billy's rock is flying through the air at t ". But then if Billy's rock had not been flying through the air at t , the bottle would still have broken when it did. So we cannot find a chain of dependence required for Billy throwing to produce the bottle to break. Since *cause* and *because* require that C produce E , we account for the unacceptability of (19).

2.4.1 Sufficiency for production

In the previous two sections we saw that *cause* and *because* require that the cause be sufficient for the effect and that the cause produce the effect. One may wonder whether these requirements interact in any way. One possibility is that the semantics of *because* requires sufficiency, and separately requires production, without any interaction between the two. We may express this as follows, where \gg denotes sufficiency and *produce* production (to avoid notational clutter we will

buffer of previous elements" we mean that for each p_t on the chain there is a time t' such that for each element $q_{t''}$ on the chain with $t' \leq t'' < t$, $\neg q_{t''}$ is sufficient for $\neg p_t$. When the chain is discrete, this condition reduces to saying that each element of the chain (excluding the first) counterfactually depends on a previous element. To say that the chain is *temporally saturated* is to say that, where p_t is the first element of the chain and $q_{t'}$ the last, for every time t'' with $t < t'' < t'$ there is a proposition r such that $r_{t''}$ is on the chain.

write everything in terms of *cause*, though one may freely replace C *cause* E with E *because* C in what follows).

Sufficiency and production. C *cause* E entails $(C \gg E) \wedge (C \text{ produce } E)$.

A stronger possibility is that *because* requires the cause to be sufficient for the *cause itself* to produce the effect:

Sufficiency for production. C *cause* E entails $C \gg (C \text{ produce } E)$.

Sufficiency for production entails sufficiency and production. This follows immediately under the following assumptions (in section 3.7 we show that they hold according to our analysis of sufficiency).

- (20) Facts showing that sufficiency for production entails sufficiency and production.
- a. **Production is factive w.r.t. the effect.** $C \text{ produce } E$ entails E .
 - b. **Cause is factive w.r.t. the cause.** $C \text{ cause } E$ entails C .
 - c. **Modus ponens for sufficiency.** $C \wedge (C \gg E)$ entails E .
 - d. **Right weakening.** If E entails E' then $C \gg E$ entails $C \gg E'$.

The converse, however, does not hold. Sufficiency for production does not entail sufficiency and production. This is because it is possible for C to be sufficient for E , while C is not sufficient for C itself to produce E . We witness the failure of this entailment in cases where the cause produced the effect, but it was possible for the cause to occur and yet for the effect to be produced in some other way. Here is a such a scenario.

Alice and Bob are two children at the funfair with their parents. The parents decide that the children should have a souvenir of their time there: if any child does not win a teddy by the end of the day, the parents will buy one for them.

Alice and Bob play a game with a spinner and a button (see Figure 2.4). A pointer moves around the circle until the player pushes the button. If the pointer lands in the thin green region, the player wins a teddy. If it lands in the red region, the player gets nothing.

Alice entered the game and, by sheer luck, pushed the spinner at the right time. The pointer landed in the winning region and she won a teddy. Bob entered the game and pushed the button at the wrong time. The pointer landed in the red region and he didn't win a teddy.

At the end of the day, the parents notice that Bob didn't win a teddy, so they bought him one.

So Alice and Bob both entered the spinner game. Alice won and Bob lost, so Alice got a teddy but Bob did not. Now consider (21).

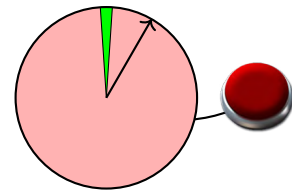


Figure 2.4: The spinner.

(21) Alice got a teddy because she entered the spinner game.

Intuitively, (21) is unacceptable. Bob is here as a foil to make salient the possibility that Alice enters the game but pushes the button at the wrong time, in which case she does not get a teddy from the spinner game.

Let us now check the following three conditions.

Sufficiency: Alice enter \gg get teddy ✓

Since Alice was guaranteed to get a teddy, anything whatsoever was sufficient for her to get a teddy. In particular, Alice entering the game was sufficient for her to get a teddy.

Production: Alice enter *produce* get teddy ✓

The fact that Alice entered the game produced her to get a teddy. Without providing a formalization of the scenario, we can loosely say that Alice entering the game produced her to get a teddy because there is a chain from her entering the game to her getting a teddy from the game such that each condition on the chain counterfactually depends on a buffer of previous conditions on the chain.

Sufficiency for production: Alice enter \gg (Alice enter *produce* get teddy) ✗

Given the set up of the scenario, it was determined that Alice would get a teddy somehow, but it was not determined how she would get it. She could have hit the button at the wrong time and lost the spinner game (a possibility emphasized by the presence of Bob), in which case she would have gotten a teddy from her parents at the end of the day rather than from winning the game. In that case her entering the game would not have produced her to get a teddy; rather, her parents buying a teddy would have produced her to get one. In all, then, the fact that Alice entered the game was not sufficient for her entering the game to produce her to win a teddy.

Thus *sufficiency for production* but not *sufficiency and production*, correctly predicts (21) to be false.

To test this explanation of (21)'s unacceptability, let us consider a different cause, one that is sufficient for the cause to produce the effect, as in (22).

(22) Alice got a teddy because she won the spinner game (and Bob got a teddy because his parents bought him one).

(22) is intuitively true. We can account for this as follows. While Alice *entering* the game was not sufficient for her entering the game to produce her to get a teddy, *winning* the game was sufficient for her winning the game to produce her to get a teddy: *Alice win* \gg *Alice win produce get teddy*. Once (21) is minimally altered to contain a cause that was sufficient for it to produce the effect, as in (22), the judgement changes. This further supports sufficiency for production.

Before moving on, let us pause to consider an alternative explanation of (21)'s falsity. One might suggest that (21) is false because even if Alice had not entered the spinner game, she would have gotten a teddy anyway (since her parents would get her a teddy if she didn't already already one by the end of the day). In the Billy and Suzy case we already saw that *because* does not require effects to counterfactually depend on their causes. But one might think there are special reasons for this, peculiar to the Billy and Suzy case. However, (22) is also true even though the effect does not counterfactually depend on the cause: if Alice hadn't won the spinner game, she would still have gotten a teddy. This suggests that the falsity of (21) is not due to the fact that if Alice had not entered the spinner game, she would have gotten a teddy anyway.

To summarise, the fact that (21) is false in the fairground scenario supports sufficiency and production: *cause* requires that the cause be sufficient for the cause itself to produce the effect.

2.5 Difference-making: the general pattern

Let us now return to the switches case from section 2.3. According to our analysis of production, the engineer pulling the lever was sufficient for it to produce the train to reach the station. We can show that pulling the lever produced the train to reach the station using a chain such as

The engineer pulls the lever at time t .
 The track is set to the right at time t' ...
 The track is set to the right at time t'' .
 The train is in position x at time t''' ...
 The train is in position y at time t'''' .
 The train reaches the station at time t''''' .

And given the set up of the scenario, the engineer pulling the lever is sufficient for the existence of such a chain. So the engineer pulling the lever is sufficient for that to produce the train to reach the station.

So we need some other way to account for the unacceptability of (17), repeated below. The challenge is to do so without requiring the effect to counterfactually depend on the cause, since we still wish to predict that (18) are fine.

- (17) a. The train reached the station because the engineer flipped the switch.
- b. The engineer flipping the switch caused the train to reach the station.
- (18) a. The bottle broke because Suzy threw her rock at it.
- b. Suzy throwing her rock at the bottle caused it to break.

Sartorio (2005) has risen to this challenge. About the Switch case, she writes:

One thing that catches the eye about Switch is that, just as the *flip* doesn't make a difference to the [train reaching the station], the *failure to flip* wouldn't have made a difference to the [train reaching the station] either. In other words, *whether or not* I flip the switch makes no difference [to the train's arrival], it only helps to determine the route that the train takes [to the station]. This suggests that what might be missing in Switch is some kind of asymmetry between my flipping the switch and my failing to flip the switch.

(Sartorio 2005:74–75)

Sartorio distills this thought into the following principle.

Sartorio's Principle. If C caused E , then, had C not occurred, the absence of C wouldn't have caused E .

Sartorio's principle represents, I believe, a major breakthrough in our understanding of causal dependence. For it provides a principled way to distinguish preemption cases (such as the Billy and Suzy case) from switching cases. Let's first see what the principle says about the Billy and Suzy case. Suzy throwing her rock caused the bottle to break. What if Suzy hadn't thrown? In that case, it is clear that Billy throwing his rock would have caused the bottle to break. What about Suzy *not* throwing? Imagine if Suzy had not thrown (in that case Billy's rock would have hit the bottle and it would have broken anyway). Consider (23) in this context.

- (23) a. Suzy not throwing her rock caused the bottle to break.
 b. The bottle broke because Suzy did not throw her rock.

These are intuitively false. This is exactly what we need for (18) to satisfy Sartorio's Principle. Sartorio's Principle is therefore compatible with the truth of (18), as desired.

In contrast, imagine for the sake of argument that the engineer flipping the switch did cause the train to reach the station. As Sartorio (2005:75) points out, both flipping the switch and not flipping the switch make the same difference with respect to the train reaching the station (determining what route it took). So if the flipping the switch caused the train to reach the station, then for the same reasons, if the engineer had not flipped the switch, that would have also caused the train to reach the station. But this violates Sartorio's Principle, so the principle correctly predicts that the engineer flipping the switch did not cause the train to reach the station.

According to Sartorio, then, causation requires a particular kind of asymmetry, which the principle makes precise. The key difference between the switch and Suzy's throw is that the former exhibits a symmetry which the latter lacks. In this way, Sartorio's Principle helps rule out trivial changes to the effect from counting as causes, such as the engineer pulling the lever.

A nice feature of Sartorio’s Principle is that it is automatically satisfied when the effect counterfactually depends on the cause (Sartorio 2005:78). This holds since causation is factive – for something to cause an effect, the effect must occur. Now suppose that the effect counterfactually depends on the cause: if the cause hadn’t been true, the effect wouldn’t have either. In that case, nothing would have caused the effect (since causation is factive). In particular, the absence of the cause wouldn’t have caused the effect, so Sartorio’s Principle is met. Thus Sartorio’s Principle, as an analysis of the difference-making idea, predicts that one way for a cause to make a difference to an effect is for the effect to counterfactually depend on the cause. Crucially, however, counterfactual dependence is not the only way to make a difference, as preemption cases show.

There is one small technical point to address before we continue. While Sartorio formulates her difference-making condition in terms of a *would*-conditional, here we will formalise it in terms of sufficiency. In section 3.2.2 we show that *would*-conditionals and sufficiency have a different modal force: *A* is sufficient for *C* just in case every *A*-world in the relevant domain is a *C*-world, while *if A, would C* is true just in case the unique selected *A*-world in that domain is a *C*-world.¹² Nonetheless, we propose in section 3.2.3 that a *would*-conditional is assertable only if its truth does not depend on which world we happen to select. The upshot is that *if A, would C* is assertable just in case *A* is sufficient for *C*. Sartorio clearly takes her principle to be assertable, so the appropriate formalisation of her principle is the following, where \Rightarrow denotes entailment and \gg sufficiency.

Sartorio’s Principle (formalised). $C \text{ cause } E \Rightarrow \neg C \gg \neg(\neg C \text{ cause } E)$

2.5.1 The Perfection Principle

As Sartorio makes clear, her principle “is not an analysis of causation. It sets a constraint on the concept of cause, and thus it helps to carve up the concept, while at the same time leaving some room for different ways of pinning it down” (Sartorio 2005:71). The compatibility of Sartorio’s principle with diverse analyses of causation is one of its greatest strengths. But to truly take advantage of this generality, we would like to be able to take any given analysis of causation and add Sartorio’s principle to it – to treat Sartorio’s principle as a separate module. This would allow any analysis whatsoever to inherit the central benefit of Sartorio’s Principle: its correct predictions in switch cases.

To turn Sartorio’s principle into such an add-on module, we need to express the principle as a necessary condition of *cause*. That is, we need to find a formula φ such that Sartorio’s principle holds just in case $C \text{ cause } E$ entails $\varphi(C, E)$ for any sentences *C* and *E*. For if we had such a formula, it would give us an automatic procedure to take any preliminary analysis of the meaning of *cause*

¹²The term ‘modal force’ comes from Kratzer’s analysis of modality (see Kratzer 1981b:45).

– call it *proto-cause* – that lacks a principled account of the difference between preemption and switch cases, and turn into an analysis that has one. We can do this by proposing:

$$C \text{ cause } E \quad \Leftrightarrow \quad C \text{ proto-cause } E \wedge \varphi(C, E).$$

Such an entry for *cause* will automatically satisfy Sartorio’s principle, since then $C \text{ cause } E$ entails $\varphi(C, E)$, and by assumption, if $C \text{ cause } E$ entails $\varphi(C, E)$, Sartorio’s principle holds. And given that Sartorio’s principle implies that $C \text{ cause } E$ entails $\varphi(C, E)$, adding $\varphi(C, E)$ as a conjunct will also not add any further entailments to our account of the meaning of *cause* beyond what Sartorio’s Principle already implies.

There is, however, a problem. Sartorio’s principle does not have the right form to tell us what this mystery formula φ is. The problem is its non-recursive nature: the principle contains *cause* in both the antecedent and consequent. In essence Sartorio’s Principle is an inequality of the form $c \geq f(c)$, where c denotes $C \text{ cause } E$, \geq represents logical strength and f takes a sentence of the form $C \text{ caused } E$ and returns $\neg C \gg \neg(\neg C \text{ cause } E)$. To turn this into our desired necessary condition, we would like to express this inequality with all the c ’s on the left hand side: $c \geq \varphi$. Given an inequality in arithmetic, say, $x \geq 3x - x^2$, in school we learn how to put the x ’s all on one side, in this case returning $x \geq 2$ (given that x is greater than 0). Unfortunately the same tricks will not work here. While the operations of arithmetic have inverses (addition/subtraction, multiplication/division), logical operations in general do not.¹³

Unfortunately, then, there is no automatic procedure to turn Sartorio’s principle into a necessary condition on *cause*. However, such a formula φ does in fact exist; indeed, it turns out to have a simple, familiar form. To see this, let us introduce some notation. For any sentences A , B and C , let $A[C/B]$ be the result of replacing every occurrence of B in A with C . For example, $((p \vee q) \wedge \neg q)[r/q] = (p \vee r) \wedge \neg r$. Let X be any sentence whatsoever, and take

$$\varphi(C, E) \quad = \quad (C \gg X) \wedge \neg(C \gg X)[\neg C/C]$$

That is, φ says that the cause is sufficient for some formula X , but this no longer holds when we negate the cause. The principle that $C \text{ cause } E$ entails some φ of this shape we will call the *Perfection Principle*.

The Perfection Principle. For any sentences C and E , there is a sentence X such that $C \text{ cause } E$ entails $C \gg X$ and $\neg(C \gg X)[\neg C/C]$.

We call this the ‘Perfection Principle’ due to its similarity with an inference pattern known as conditional perfection. Geis and Zwicky (1971) observe that an utterance of *if* A , B “invites the inference” of *if* $\neg A$, $\neg B$. For example,

¹³Though there is a rich literature on the topic of logical subtraction, such as Peirce (1867), Jaeger (1973, 1976), Hudson (1975), Fuhrmann (1996, 1999), Humberstone (1981, 2011), Yablo (2014), and Hoek (2018).

- (24) If you mow the lawn I'll give you five dollars.

What if you don't mow the lawn? Will you get five dollars? An utterance of (24) intuitively suggests that you won't.¹⁴ This inference is not part of the literal meaning of the conditional. For instance, in many contexts it is defeasible:¹⁵

- (25) If you mow the lawn I'll give you five dollars. Also if you clean your room I'll give you five dollars.
 \nrightarrow If you don't mow the lawn I won't give you five dollars.

We see the same pattern with *would*-conditionals. For example,

- (26) If Andrew were here, Barbara would be happy.
 \leadsto Barbara is not happy.

This inference is predicted given conditional perfection, for then (26) implies that if Andrew were not here, Barbara would not be happy. This, together with the inference that Andrew is in fact not here, implies that Barbara is not happy.¹⁶

Geis and Zwicky write that conditional perfection is the inference *if* $\neg A$, $\neg B$. Following Stalnaker (1968), Mandelkern (2018), Cariani and Santorio (2018) and others, we argue in section 3.2.2 that *will* and *would* select a single world at which to evaluate the consequent. This renders *if* $\neg A$, $\neg B$ equivalent to $\neg(\text{if } \neg A, B)$ for *will* and *would* conditionals. Assuming that A is not a subsentence of B , then we can express conditional perfection with *will/would*-conditionals as the inference from *if* A, B to $\neg(\text{if } A, B)[\neg A/A]$, which fits the pattern of the Perfection Principle.

Now for our main result of this section. Under plausible assumptions about \gg , Santorio's Principle is equivalent to the Perfection Principle. Those assumptions are given in (27), where $A \diamond \rightarrow C$ abbreviates $\neg(A \gg \neg C)$.

- (27) a. **Nonempty domains.** $A \gg C$ entails $A \diamond \rightarrow C$.
 b. **Stability.** $C \text{ cause } E$ entails $C \gg (C \text{ cause } E)$.
 c. **Idempotence.** $A \diamond \rightarrow C$ entails $A \gg (A \diamond \rightarrow C)$.
 d. **Right weakening.** If C entails C' then $A \gg C$ entails $A \gg C'$.
 e. If $C \text{ cause } E$ is true, then C is not a subsentence of E .

(27a) says that the set of worlds we consider when we evaluate sufficiency is nonempty: if every relevant A -world is a C -world, some relevant A -world is a C -

¹⁴For an overview of the pragmatic principles that may underlie conditional perfection, see von Fintel (2001a). Nadathur and Lauer (2020:§4.2) propose that an inference analogous to conditional perfection can also occur with causal words such as *make*.

¹⁵For a discussion of cases where conditional perfection is indefeasible, see Van Canegem-Ardijns (2010).

¹⁶For discussion of the inference from a *would*-conditional to the falsity of the antecedent, see Ippolito (2003), Starr (2014), and Leahy (2011, 2018). For discussion of the inference that the consequent is false, see Karttunen (1971) and Ross (2020).

world. This is a standard assumption to make – common to all quantificational elements (Cooper 1983, von Fintel 1994, Beaver 1995, Ippolito 2006). (27b) is a stability principle. It says that if C caused E , then C was sufficient for it to cause E . (27c) says that conditional restriction is idempotent: if when we restrict to the relevant A -worlds, we find a C -world, then restricting to the relevant A -worlds, and then restricting to the A -worlds again, we still find a C -world. Lastly, (27d) says that sufficiency satisfies right-weakening. (27b)–(27d) are satisfied by our analysis of sufficiency and the semantics of *cause* we will ultimately propose.

These assumptions are also plausible for other constructions besides sufficiency. For example, if we replace \gg with a *would*-conditional ($>$), then stability follows from conjunctive sufficiency – the inference from $A \wedge C$ to $A > C$ – together with the factivity of *cause*. By factivity, $C \text{ cause } E$ entails $C \wedge (C \text{ cause } E)$, which by conjunctive sufficiency entails $C > (C \text{ cause } E)$.

2.5.1. THEOREM. *Sartorio's Principle is equivalent to the Perfection Principle, given the assumptions in (27).*

PROOF. (\Rightarrow) Suppose Sartorio's Principle. Pick any sentences C and E and take $X = (C \text{ cause } E)$. Then by Stability, $C \text{ cause } E$ entails $C \gg (C \text{ cause } E)$, which is $C \gg X$. We also have the following chain of implications.

$$\begin{array}{ll}
C \text{ cause } E & \\
\neg C \gg \neg(C \text{ cause } E) & \text{(Sartorio's Principle)} \\
\neg(\neg C \gg \neg(C \text{ cause } E)) & \text{(Nonempty domains)} \\
\neg(C \gg (C \text{ cause } E))[\neg C/C] & (C \text{ does not appear in } E) \\
\neg(C \gg X)[\neg C/C] & (X = C \text{ cause } E)
\end{array}$$

Hence $C \text{ cause } E$ entails $C \gg X$ and $\neg(C \gg X)[\neg C/C]$.

(\Leftarrow) Suppose the Perfection Principle. So $\neg C \text{ cause } E$ entails $(C \gg X)[\neg C/C]$. Then by contraposition we have (\dagger): $\neg(C \gg X)[\neg C/C]$ entails $\neg(\neg C \text{ cause } E)$. Observe the following chain of implications.

$$\begin{array}{ll}
C \text{ cause } E & \\
\neg(C \gg X)[\neg C/C] & \text{(Perfection Principle)} \\
\neg(\neg C \gg X[\neg C/C]) & \text{(Definition of } [\neg C/C]) \\
\neg C \diamond \rightarrow \neg X[\neg C/C] & \text{(Definition of } \diamond \rightarrow) \\
\neg C \gg (\neg C \diamond \rightarrow \neg X[\neg C/C]) & \text{(Idempotence)} \\
\neg C \gg \neg(\neg C \gg X[\neg C/C]) & \text{(Definition of } \diamond \rightarrow) \\
\neg C \gg \neg(C \gg X)[\neg C/C] & \text{(Definition of } [\neg C/C]) \\
\neg C \gg \neg(\neg C \text{ cause } E) & (\dagger \text{ and right weakening)}
\end{array}$$

Hence $C \text{ cause } E$ entails $\neg C \gg \neg(\neg C \text{ cause } E)$, which is Sartorio's Principle. \square

2.5.2 The ubiquity of the Perfection Principle

When we examine the literature on causation, we see Perfection Principle show up time and time again. A number of proposals exhibit the asymmetry between the cause and its absence that the principle demands. Here are some examples.

- Lewis (1973a:536) proposes that an event e causally depends on an event c just in case the following two counterfactuals are true: if c had occurred, e would have occurred, and if c had not occurred, e would not have occurred; in symbols, $O(c) \Box \rightarrow O(e)$ and $\neg O(c) \Box \rightarrow \neg O(e)$. Assuming that the cause not occurring is a counterfactual possibility, this implies $\neg(\neg O(c) \Box \rightarrow O(e))$, which gives the asymmetry required by the Perfection Principle.
- Wright (1985, 2011) proposes the NESS (Necessary Element of a Sufficient Set) test for causation, according to which something is a cause just in case there is a set of conditions that are jointly sufficient for the effect, but are not sufficient when the cause is removed from the set.
- Mackie’s INUS condition states that a cause is “an insufficient but non-redundant part of a condition which is itself unnecessary but sufficient for the result” (Mackie 1974:64), which implies that some condition $B \cup \{C\}$ is sufficient for E but B is not. Assuming that Mackie’s notion of sufficiency expresses universal quantification over some set of possibilities, to say that B is not sufficient for E is to say that E is false at some relevant B -possibility. Given that $B \cup \{C\}$ is sufficient for E , this possibility must be a $\neg C$ -possibility, so Mackie’s INUS condition implies that $B \cup \{C\}$ is sufficient for E but $B \cup \{\neg C\}$ is not – as the Perfection Principle requires.
- More recently, Beckers (2016) use a notion of *production*, arguing that the semantics of *is a cause of* involves comparing the presence and absence of the cause with respect to producing the effect. According to Beckers, C is an cause of E just in case, informally put, C produced E , and after intervening to make $\neg C$ true, $\neg C$ would not have also produced E ; in symbols $C \text{ produce } E \wedge [\neg C = 1](\neg C \text{ produce } E)$. Given the modelling framework Beckers uses (structural causal models), this is equivalent to $[C = 1](C \text{ produce } E) \wedge [\neg C = 1](\neg C \text{ produce } E)$, which fits the shape required by the Perfection Principle.

Despite the many differences between these accounts, the Perfection Principle emerges as a common thread throughout the history of work on causation.

2.5.3 On the pragmatic origins of the Perfection Principle

In light of the parallels between conditional perfection and the Perfection Principle, it is natural to wonder whether the Perfection Principle could have pragmatic origins. In their paper on conditional perfection, Geis and Zwicky write,

Certainly, it seems to be the case that an inference can, historically, become part of semantic representation in the strict sense; thus, the development of the English conjunction *since* from a purely temporal word to a marker of causation can be interpreted as a change from a principle of invited inference associated with *since* (by virtue of its temporal meaning) to a piece of the semantic content of *since*.

(Geis and Zwicky 1971:565–566)

To illustrate Geis and Zwicky’s observation, consider:

- (28) Maher has been out of town since Sue told him that she loves him.

This is ambiguous between a temporal and causal reading of *since*. Maher might love Sue too but simply be away on a work trip (temporal *since*), in which case if Sue hadn’t confessed her feelings he would be out of town anyway. Or he might be out of town because Sue told him that she loves him (causal *since*).

Similarly, the word *then* has a temporal meaning (*Tamara yawned then Omri yawned*). But Iatridou (1993, 2021) observes that *then* in conditionals takes on a further meaning. She offers the following examples, which are unacceptable with *then* but fine without it.

- (29) a. If I may be frank (*then) you are not looking good today.
 b. If John is dead or alive (*then) Bill will find him.
 c. If he were the last man on earth (*then) she wouldn’t marry him.
 d. Even if you give me a million dollars (*then) I will not sell you my piano.

Where $O(p)(q)$ denotes the conditional construction, Iatridou (1993) proposes that *if p then q* asserts $O(p)(q)$ and presupposes $\neg O(\neg p)(q)$. (The similarity between Iatridou’s proposal and conditional perfection is unmistakable.) Iatridou’s proposal accounts for the fact that *then* is unacceptable in the sentences in (29), since they are incompatible with the existence of relevant $\neg p \wedge \neg q$ cases. For example, (29b) would require a case in which John is neither dead nor alive.

We are starting to see a pattern. Natural language has a habit of taking a temporal/conditional meaning and adding its perfection to its semantic content. It is tempting to think that something similar might have happened to *cause* and *because*. Of course, this is just a suggestive remark – detailed diachronic investigations would be needed to test this claim. But the pattern is striking.

2.5.4 Adding the Perfection Principle

Theorem 2.5.1 gives us a straightforward way for our account to satisfy Sartorio’s Principle, and thereby inherit its account of the switches case. We have already seen that *cause* and *because* entail that the cause is sufficient to produce the effect:

$C \gg (C \text{ produce } E)$. Looking at the Perfection Principle, we can simply take X to be $C \text{ produce } E$. To satisfy Sartorio's Principle we only have to add that $C \gg (C \text{ produce } E)$ does not hold when we replace the cause with its negation: $\neg(\neg C \gg (\neg C \text{ produce } E))$. The absence of the cause is not sufficient for it to produce the effect.¹⁷

Adding this condition to sufficiency for production gives us the following formula, which we propose is the semantics of *cause* and *because*. For any sentences C and E , $C \text{ cause } E$ and $E \text{ because } C$ are true just in case the following is true.

$$C \wedge C \gg (C \text{ produce } E) \wedge \neg(\neg C \gg (\neg C \text{ produce } E))$$

A glance at this semantics reveals that it has the right shape to satisfy the Perfection Principle, and so by Theorem 2.5.1 satisfies Sartorio's principle. It therefore automatically inherits Sartorio's account of the switches case. Nonetheless, it can be instructive to show this directly. On our analysis of production, if the engineer hadn't pulled the lever, that would have also produced the train to reach the station. We can show this via the chain:

The engineer does not pull the lever at time t .
 The track is set to the left at time t' ...
 The track is set to the left at time t'' .
 The train is moving at position x' at time t''' ...
 The train is moving at position y' at time t'''' .
 The train reaches the station at time t''''' .

And given the set up the scenario (if the engineer hadn't pulled the lever, the train was guaranteed to take the left track), not pulling the lever is sufficient for there to exist a chain of the kind of above. So not pulling the lever is sufficient for that to produce the train to reach the station: $\neg C \gg (\neg C \text{ produce } E)$. Our semantics therefore correctly predicts (17) to be false.

At the same time, we preserve our account of the Billy and Suzy case. Suzy not throwing her rock is not sufficient to produce the bottle to break. This is intuitively plausible: if we imagine Suzy not throwing her rock, we do not find any chain of dependence from Suzy not throwing to the bottle breaking (whereas in the switches case we do find a chain of dependence from the engineer not pulling the lever to the train reaching the station). One reason for this is that, since Billy threw after Suzy, we may consider it possible that if Suzy had not thrown, Billy might not have thrown at all, or might have thrown inaccurately, in which case the bottle might not have broken.¹⁸ Thus Suzy not throwing is not sufficient

¹⁷This is exactly the condition Beckers (2016:93) proposes to account for the switching case. Beckers formulates his account in terms of structural causal models, representing the conditional construction in terms of interventions, and proposes if C is an actual cause of E , then under an intervention $\neg C$, $\neg C$ would not have produced E .

¹⁸Beckers (2016:86) makes this point for a slight variant of the Billy and Suzy case.

for the bottle to break. It follows that Suzy not throwing is not sufficient for that to produce the bottle to break. Since Suzy actually threw the rock and that was sufficient to produce the bottle to break, our semantics of *cause* and *because* predicts (18) to be true.¹⁹

This raises the question whether *cause* and *because* require that, if the cause had not occurred, the effect might not have occurred. Unlike many other proposals, the present account does not require this.²⁰ It is possible for *C cause E* and *E because C* to be true even when *E*'s truth was inevitable, in the sense that in every nomically possible world where *C* becomes true, *E* eventually becomes true too. Consider, for example:

- (30) a. Socrates drinking poison caused him to die.
b. Socrates died because he drank poison.

These are straightforwardly true. Given the inevitability of death, this appears to be a problem for approaches that appeal to some possibility where, if the cause had not occurred, the effect would not have occurred. There have been many responses to this problem on behalf of such approaches; for example, that the relata of causal claims are fine-grained events, or that we restrict attention to a salient period of time after the cause occurred. Section 5.6 discusses these responses in detail and gives a number of arguments against them.

In contrast, the present approach accounts for the truth of (30) right out of the box, without special assumptions. We do not require that if Socrates hadn't drunk poison, he might not have died. Instead, we merely require that if Socrates hadn't drunk poison, him not drinking poison might not have produced him to die. On our analysis of production, this means that it is possible for there to be no chain of counterfactual dependence from him not drinking poison at some time to being dead at some time. While it is hard to give a general proof of this claim (and we will not do so here), we can consider some plausible chains and see how they fail. Here are three attempts.

S isn't drinking poison at t	S isn't drinking poison at t	S isn't drinking poison at t
S is alive at t' ...	S's body is in condition x at t' ...	The world is in state x at t'
S is alive at t''	S's body is in condition y at t''	The world is in state y at t'
S is dead at t'''	Socrates is dead at t'''	S is dead at t'''

Each fails to be a chain of counterfactual dependence. For the first, Socrates

¹⁹This follows from factivity of production and right weakening of $\Diamond \rightarrow$. By factivity, $\neg C$ produce E entails E . So $\neg E$ entails $\neg(\neg C$ produce E). Then by right weakening, $\neg C \Diamond \rightarrow \neg E$ entails $\neg C \Diamond \rightarrow \neg(\neg C$ produce E), i.e. $\neg(\neg C \gg E)$ entails $\neg(\neg C \gg (\neg C$ produce E)). That is, if $\neg C$ is not sufficient for E then $\neg C$ is not sufficient for it to produce E . For further discussion of the semantics of $\Diamond \rightarrow$ see section 4.1.2.

²⁰Proposals that appeal to a possibility where the effect does not occur include Lewis (1973a), Yablo (2004), Halpern and Pearl (2005), Weslake (2015), Halpern (2016), Beckers and Vennekens (2018), Beckers (2021a), and Andreas and Günther (2020, 2021).

not being alive at t'' is not sufficient for him to not be dead at t''' . Similarly, Socrates's body not being in condition y at t'' is not sufficient for him to be dead at t''' . This is because sufficiency requires, loosely speaking, for us to consider *all* the ways for Socrates' body to not be in condition y . Presumably, for some of these ways, Socrates would still have died at t''' . The third chain does not count as a chain of counterfactual dependence for the same reason. There are many ways for the world to not be in the particular state it is in at t' . For some of these, Socrates would still have died at t''' .

In contrast, Socrates drinking poison is sufficient to produce him to die. That is, drinking poison is sufficient for there to be a chain of counterfactual dependence from him drinking poison to being dead; such as the following (apologies for the morbid detail).

Socrates is drinking poison at t
 The poison is travelling through Socrates' body at t' ...
 Socrates stops breathing at t'' ...
 Socrates is dead at t'''

So Socrates drank poison, him drinking poison was sufficient to produce him to die, but not drinking poison was not. Our semantics therefore correctly (30) to be true, as desired.

2.6 Cause = difference-making + sufficiency + production

In this brief section I would like to pause to appreciate that the semantics of *cause* and *because* we propose has an especially simple structure. We can express our proposal using three relations between sentences: sufficiency, production and difference-making. We formalise sufficiency in Chapter 3 and production in Chapter 5. In light of the Perfection Principle, let us formalise the difference-making relation as follows.

(31) A makes a difference to B just in case $A \wedge B \wedge \neg B[\neg A/A]$ is true.

This is a factive notion of difference-making: for A to make a difference to B , they must both be true. Difference-making adds that when we replace A with its negation in B , the result is no longer true. This is a very simple, syntactic implementation of the difference-making idea, one that will allow us to express our semantics of *cause* and *because* economically.

Given the three relations we have discussed – sufficiency, production and difference-making – we can combine into a single relation in a straightforward way. Consider the following way to construct new binary relations between sentences from old ones. For any relations R and S between sentences, let us define

the relation $R + S$ given by

$$\varphi(R + S)\psi \quad \text{just in case} \quad \varphi R(\varphi S\psi)$$

for any sentences φ and ψ . That is, φ is $(R + S)$ -related to ψ just in case φ is R -related to the sentence that φ is S -related to ψ . The $+$ operation applies the relations in order while copying the left argument each time.²¹ This $+$ operation may represent, for example, making a restriction to φ -worlds and then having every subsequent operation inherit that restriction.

With this operation, we can express our semantics of *cause* very simply:

$$\text{cause} = \text{difference-making} + \text{sufficiency} + \text{production}$$

as shown by the following derivation.²²

$$\begin{aligned} C \text{ cause } E &= C \text{ DM} + \gg + \text{produce } E \\ &= C \text{ DM } (C \gg + \text{produce } E) \\ &= C \text{ DM } ((C \gg (C \text{ produce } E)) \\ &= C \wedge C \gg (C \text{ produce } E) \wedge (C \gg (C \text{ produce } E)[\neg C/C] \\ &= C \wedge C \gg (C \text{ produce } E) \wedge (\neg C \gg (\neg C \text{ produce } E)) \end{aligned}$$

In a sentence, *cause* means difference-making sufficiency for production.

Here is yet another way to express our semantics. First, compare:

- (32) a. Alice wants Bill to sing.
b. Alice wants to sing.

To treat these two constructions in a uniform way, among other reasons, syntacticians often assume the existence of a silent determiner phrase, called PRO (see Sportiche, Koopman, and Stabler 2013:240ff.), as in (33).

- (33) Alice_i wants PRO_i to sing.

This can be paraphrased as *Alice wants Alice to sing*.²³ Similarly, we can express our semantics of *cause* as:

²¹This operation turns out to have a familiar mathematical structure: for any set of sentences L closed under $+$, $(L, +, 0)$ is a monoid, with the identity element given by $\varphi 0 \psi = \psi$. This follows since $+$ is associative: $R + (S + T) = (R + S) + T$ for any relations R, S, T .

²²Here we have unpacked $\text{DM} + \gg + \text{produce}$ as $\text{DM} + (\gg + \text{produce})$, though since $+$ is associative, this is equivalent to unpacking it via $(\text{DM} + \gg) + \text{produce}$.

²³Indeed, we can add PROs indefinitely, as shown in (i).

- (i) Alice decided to try to visit to museum to ask to see the manuscript.

whose meaning can be paraphrased as: Alice decided that Alice should try to have Alice visit the museum so that Alice can ask for Alice to see the manuscript.

- (34) C makes a difference to being sufficient to produce E .

since this is interpreted as

- (35) a. C_i makes a difference to PRO_i being sufficient to PRO_i produce E .
 b. C makes a difference to C being sufficient to C produce E .

2.7 The modal force of difference-making

We have seen evidence that *cause* and *because* (i) entail that the cause is sufficient to produce the effect and (ii) satisfy the Perfection Principle. Our proposed semantics takes the most direct route available to satisfy both: we simply add that the cause's negation is not sufficient to produce the effect. Given that sufficiency is a universal claim, negating it results in an existential claim: in *some* relevant world where the cause does not occur, its absence does not produce the effect.²⁴

Though looking again at the Perfection Principle, we see that there are in fact many routes to it. For instance, we could instead propose that *cause* and *because* entail that the cause's negation *is* sufficient to *not* produce the effect; that is, in every relevant where the cause does not occur, its absence does not produce the effect. We can express the difference in terms of the scope of negation.

- (36) C *cause* E and E *because* C entail ...
 a. Existential difference-making: $\neg(\neg C \gg (\neg C \text{ produce } E))$
 b. Universal difference-making: $\neg C \gg \neg(\neg C \text{ produce } E)$

Assuming that there is some relevant world where the cause does not occur, the universal condition implies the existential condition we propose. To satisfy the Perfection Principle, the universal condition is gratuitous. But could it be right?

Consider the following sentences, discussed by McHugh (2020), found in newspapers and websites.²⁵

- (37) a. Reyna was born at Royal Bolton Hospital but received a Danish passport because her mother was born in Copenhagen.
 b. He has an American passport because he was born in Boston.
 c. I think I was laid off because I'm 56 years old.
 d. Naama Issachar ... could spend up to seven-and-a-half years in a Russian prison because 9.5 grams of cannabis were found in her possession during a routine security check.

²⁴For a formalisation of what it means to be a world to be 'relevant' in this sense, see our analysis of sufficiency in section 3.6.

²⁵Sources: (37a) [The Bolton News](#), 12 February 2020; (37b) [RuPaulsDragRace.fandom.com](#); (37c) [The Chicago Tribune](#), 7 September 2003; (37d) [The Jerusalem Post](#), 24 November 2019; (37e) [Healthline.com](#), 13 June 2019.

- e. A 90-day study in 8 adults found that supplementing a standard diet with 1.3 cups (100 grams) of fresh coconut daily caused significant weight loss.

Consider (37a). If Reyna’s mother hadn’t been born in Copenhagen, there are intuitively many places where she could have been born instead – some inside Denmark (Elsinore, Køge, ...), others outside. In some of these cases Reyna would still have received a Danish passport, in others not. In other words, the following *would*-conditional is unacceptable.

- (38) If Reyna’s mother hadn’t been born in Copenhagen, Reyna wouldn’t have received a Danish passport.

In contrast, the *because* claim in (37a) sounds fine.

Our existential difference-making condition correctly predicts this. It is enough to observe that, if Reyna’s mother had instead been born outside Copenhagen, among the possibilities is one where she is born outside Denmark and Reyna does not receive a Danish passport: $\neg C$ is not sufficient for E , so (by factivity of production and right weakening) $\neg C$ is not sufficient for $\neg C$ produce E . The existential difference-making condition is satisfied.

However, the universal difference-making condition, $\neg C \gg \neg(\neg C \text{ produce } E)$ incorrectly predicts (37a) to be false. This is because, if Reyna’s mother hadn’t been born in Copenhagen, among the possibilities we consider is one where she is still born in Denmark; say, in Aarhus. In that case our analysis of production predicts that Reyna not being born in Denmark produces Reyna to get a Danish passport: there is a chain of dependence from Reyna not being born in Copenhagen to receiving a Danish passport. (Without going into too many details about production here, this is because in a world where her mother is born in Aarhus, “Reyna’s mother was not born in Denmark” and “Reyna’s mother was born in Aarhus” are about the same state.)

The fact that (37) are acceptable is evidence that *cause* and *because* allow the cause to be stronger than strictly required for the claim to be true. Is there any limit on how strong the cause can be? Consider the following naturally-occurring examples (also discussed by McHugh 2020).²⁶

- (39) a. Computers do an awful lot of deliberation, and yet their every decision is wholly caused by the state of the universe plus the laws of nature.
- b. If anything is happening at this moment in time, it is completely dependent on, or caused by, the state of the universe, as the most complete description, at the previous moment.

²⁶Sources: (39a) CommonsenseAtheism.com; (39b) CausalConsciousness.com; (39c) Edge.com.

- c. If you keep asking “why” questions about what happens in the universe, you ultimately reach the answer “because of the state of the universe and the laws of nature.”

The status of these sentences is arguably more controversial than those in (37). Nonetheless, their authors clearly take them to be assertable. This suggests that a *cause* or *because* sentence can be assertable even when the cause is far stronger than required for the claim to be true.

One last datum before we move on. Figure 2.5, from the internet, illustrates the absurdity of a universal difference-making condition. *Swissguy25*’s response is an instance of *Post hoc ergo propter hoc* – inferring that since the man was paralysed *after* he ate 413 chicken nuggets, he must have been paralysed *because* he ate 413 chicken nuggets. Nonetheless, in this case that is a perfectly reasonable conclusion. But there is another fallacy at play: taking causation to have a universal difference-making condition. Consider what happens if we try to paraphrase this causal claim with a *would*-conditional.



Figure 2.5: [Memezila.com](https://www.memezila.com)

- (40) If he hadn’t eaten 413 chicken nuggets, he wouldn’t have been paralysed.

This seems to license *swissguy25*’s conclusion that 412 chicken nuggets is a safe portion size. Proposing a universal difference-making condition not only leads to the wrong predictions. It is also a public health hazard.

2.7.1 The implicatures of difference-making

One reason why a universal difference-making condition may seem plausible comes from cases of irrelevant causes. Suppose that Alice flicked a switch and Bob sneezed. The light turned on. Consider (41).²⁷

- (41) a. Alice flicking the switch and Bob sneezing caused the light to turn on.
b. The light turned on because Alice flicked the switch and Bob sneezed.

These are intuitively unacceptable. But if it were false that Alice flicked the

²⁷Similar examples can be found in discussions of whether causes must be proportional to their effects, such as the red/scarlet example from Yablo (1992). For discussions of proportionality see Bontly (2005), Shapiro and Sober (2012), Weslake (2013), and McDonnell (2017), among others. A closely related issue is whether causation is a contrastive relation (see Schaffer 2005, 2010). For a compelling account of the source of contrastivity in causal claims, see Rooth (1999) and Beaver and Clark (2008:63–66).

switch and Bob sneezed, there are many possibilities to consider. In some of these, Alice doesn't flick the switch and the light stays off (for an analysis of how logically complex sentences raised hypothetical scenarios, see chapter 4). Also, Alice flicking the switch and Bob sneezing was sufficient to produce the light to turn on. So (41) are true according to the existential difference-making condition but false according to the universal difference-making condition. Is this not evidence in favour of the universal condition?

McHugh (2020) argues that (41) are unacceptable due to a false implicature. Notice that the causes in (41) have a different structure than those in (37) and (39): the causes in (41) take the form of a conjunction, while the causes in (37) and (39) do not. It is commonly agreed that implicatures are sensitive to which alternative utterances are available.²⁸ Katzir (2007) and Fox and Katzir (2011) propose that for any sentence *S*, the sentences that result from deleting material from *S* are alternatives to *S*, what Katzir calls its *deletion alternatives*. This makes intuitive sense. If someone makes an utterance, we would expect there to be a reason for uttering each part of it. Why waste your breath? Following this work, we propose that the sentences in (41) have the following alternatives, respectively.

- (42) a. Alice flicking the switch caused the light to turn on.
 b. The light turned on because Alice flicked the switch.

A further assumption often made is that when hearers interpret a sentence *S*, they assume by default that every alternative to *S* that *S* does not entail is false (see e.g. van Rooij and Schulz 2007). The sentences in (41) do not entail their alternatives in (42). We therefore expect (41) to trigger the implicature that (42) are false. Given that these alternatives are in fact true, the implicature is false, which accounts for the unacceptability of (41).

An immediate prediction of this account is that, given a causal claim where the cause is stronger than required, its acceptability depends on which alternatives are available. This prediction appears to be borne out. Following the theory of alternative calculation from Fox and Katzir (2011), we can make *Denmark* an alternative by making it contextually salient and focusing *Copenhagen*, as in the following dialogue.

- (43) A: I have a Danish passport because my father was born in Denmark.
 Why do you have one?
 B: Because my mother was born in COPENHAGEN.

As McHugh (2020:134) observes, B's utterance suggests that Copenhagen is somehow special when it comes to receiving Danish passports; in other words, that it is not true that B has a Danish passport because their mother was born in Den-

²⁸For some recent discussion of this point see Van Tiel and Schaeken (2017), Repp and Spalek (2021), Bott and Frisson (2022), and Zhang et al. (2023).

mark. Thus the difference between (37a), which is assertable, and (41), which are not, is a difference in the availability of alternatives. (37a) is assertable when *Denmark* is not an alternative to *Copenhagen*, but (41) is always unassertable because deletion alternatives are always active.

2.7.2 *Only because*

A popular idea is that there is a deep connection between implicatures and the meaning of *only* (van Rooij and Schulz 2004, Schulz and van Rooij 2006, Spector 2003, 2007, Fox 2007, Fox and Spector 2018). Implicatures can often be paraphrased with *only*. Consider:

- (44) Reyna received a Danish passport *only because* her mother was born in Copenhagen.

This has a number of readings. This is expected, since it is well-known that *only* is alternative-sensitive (Rooth 1985). Different alternative sets give rise to different readings. We adopt the following standard entry of *only*, which incorporates alternative-sensitivity.²⁹

- (45) **Meaning of *only*.** For any sentence *S* and set of sentences *Alt*, *only_{Alt} S* asserts that for every $A \in Alt$, if *S* does not entail *A* then *A* is false.

To illustrate with the classic example, compare:

- (46) a. I *only* introduced BILL to Sue.
b. I *only* introduced Bill to SUE.

In (46a), *only* negates alternatives of the form *I introduced x to Sue*, saying I didn't introduce anyone but Bill to Sue, while in (46b) it negates alternatives of the form *I introduced Bill to x*, saying I didn't introduce Bill to anyone but Sue.

To illustrate some of the readings of (44), suppose both of Reyna's parents were born in Copenhagen, but in Reyna's case the law *only* allows her mother, not her father, to pass on citizenship to her. In that case (44) may have a single alternative:

- (47) Reyna received a Danish passport *because* her father was born in Copenhagen.

This alternative can be triggered by focus on *mother* in (44). Given this alternative, (44) implies that (47) is false. If the law *only* allows Reyna's mother to pass on citizenship to her, (47) is indeed false, so on this reading (44) is fine.

Now, (44) has another reading on which it is intuitively false. On this reading, (44) implies that if Reyna's mother hadn't been born in Copenhagen, Reyna

²⁹Our predictions in this section also follow on Fox's (2007) entry for *only*.

wouldn't have received a Danish passport.

How does the present account of the meaning of *because* predict this reading?

We propose that when *only because* has a counterfactual dependence reading, the alternatives are all other *because*-clauses; that is, the alternatives to *E only because C* are sentences of the form *E because D* where *D* is a sentence. This appears to be the most generic, hands-off choice of alternatives to assume.³⁰ Then among the alternatives to (44), we have:

- (48) Reyna received a Danish passport because her mother was born in Denmark.

The prejacent of *only* in (44), *Reyna received a Danish passport because her mother was born in Denmark*, does not entail this alternative. For example, there is a logically possible world where only those born in Copenhagen receive Danish passports. In that world (44) is true but (48) is false since it fails the sufficiency requirement of *because*. Thus *only* negates it. But (48) is actually true. (44) entails something false, which accounts for its unacceptability.

We still have to derive the counterfactual dependence inference, that if Reyna's mother hadn't been born in Copenhagen, Reyna wouldn't have received a Danish passport. Given our assumption about the set of alternatives, for any sentence *B*, *E because (C ∨ B)* is an alternative to *E only because C*. Now let *X* be any place in Denmark, and consider the sentence, which we abbreviate as *E because (C ∨ X)*.

- (49) Reyna received a Danish passport because her mother was born in Copenhagen or *X*.

If *X* is included in Copenhagen, then *E because C* entails *E only because (C ∨ X)*, so *E only because C* does not assert $\neg(E \text{ because } (C \vee X))$. But for any *X* not included in Copenhagen, *E because C* does not entail *E because (C ∨ X)*. (This is for the same reason as above: there is a logically possible world where only those with parents born in Copenhagen receive Danish passports.) So *E only because C* asserts $\neg(E \text{ because } (C \vee X))$. On our semantics of *because* this means

$$\begin{aligned} & \neg(C \vee X) \\ \vee & \neg((C \vee X) \gg ((C \vee X) \text{ produce } E)) \\ \vee & (\neg(C \vee X) \gg (\neg(C \vee X) \text{ produce } E)) \end{aligned}$$

The first disjunct is false: Reyna's mother was born in Copenhagen, and so born in Copenhagen or *X*. The third disjunct is also false. It says that her mother being born outside Copenhagen and *X* is sufficient to produce Reyna to

³⁰There is independent evidence for this choice of alternatives from other environments. For example, von Stechow (1997:28, taking up an idea by Roger Schwarzschild) and Vostrikova (2018) suggest that *only if* has as alternatives the set of all *if*-clauses.

receive a Danish passport. If she had born outside Copenhagen and X , there are possibilities in which she is born outside Denmark altogether, in which case Reyna does not receive a Danish passport, and so nothing produces Reyna to receive a Danish passport.

That leaves the second disjunct. It says that if Reyna's mother had born in Copenhagen or X , there is a possibility in which this did not produce Reyna to receive a Danish passport. Plausibly, we may assume that this is not a preemption case: there is no backup which would produce Reyna to get a Danish passport apart from through her mother. Thus if Reyna received a Danish passport, her mother being born where she was produced her to receive one:

$$(C \vee X) \gg E \quad \text{implies} \quad (C \vee X) \gg ((C \vee X) \text{ produce } E).$$

Then by contraposition,

$$\neg((C \vee X) \gg (C \vee X) \text{ produce } E) \quad \text{implies} \quad \neg((C \vee X) \gg E).$$

So Reyna's mother being born in Copenhagen or X is not sufficient for her to receive a Danish passport. This argument holds for every place X outside Copenhagen. *E only because C* also implies *E because C*.³¹ So every relevant C -case is an E -case, but for every X not implied by C , some relevant $(C \vee X)$ -case is not an E -case. This implies that every relevant $\neg C$ -case is a $\neg E$ -case. We therefore predict the counterfactual dependence inference.

While we have focused on a specific case, the reasoning above applies broadly. Assuming we are not in a preemption case, *only because* asserts, in essence, that there are no backup causes. For any backup X , $C \vee X$ is not sufficient for E . This, together with the inference that C is sufficient for E , implies that E counterfactually depends on C .

2.7.3 Why counterfactual dependence is so compelling

I would like to end this chapter by considering one case in particular where the counterfactual dependence inference of *only because* has played a major role. Imagine David Hume, sitting in his armchair and pondering causation. He imagines hitting a billiard ball. Hitting the ball caused it to move. Why is that true? He thinks to himself: *well, if I hadn't hit the ball, it wouldn't have moved*. And so he formulates a hypothesis:

- (50) "Hitting the ball caused it to move" is true because if I hadn't hit the ball, it wouldn't have moved.

It is easy to take for granted that hitting the ball is sufficient for it to move (indeed, that hitting the ball is sufficient for that to produce it to move) – that is clear, it

³¹There has been much discussion about the nature of this implication (see Ippolito 2008). Nonetheless, it seems intuitively clear in this case that (44) implies (37a).

goes without saying. Given this assumption, the whole *because* sentence in (50) is true. Indeed, our analysis of *because* predicts it to be true, since counterfactual dependence implies the difference-making condition we propose (given that the cause not occurring is a counterfactual possibility).

$$\neg C \gg \neg E \quad \Rightarrow \quad \neg C \gg \neg(\neg C \text{ produce } E) \quad \Rightarrow \quad \neg(\neg C \gg (\neg C \text{ produce } E))$$

So (50) is yet another example of a *because* sentence that is true even though the cause is stronger than required – just like the sentences in (37).

Now, as Groenendijk and Stokhof (1984) observe, we have a pervasive tendency to interpret answers to questions exhaustively. For example:

- (51) A: Who did you have lunch with?
B: Katrin and Maria.

B's answer intuitively suggests that B *only* had lunch with Katrin and Maria. Similarly, given the question "Why is it true that hitting the billiard ball caused it to move?" it is perfectly natural to interpret the answer in (50) exhaustively. Given this exhaustive interpretation, Hume therefore comes to believe:

- (52) "Hitting the ball caused it to move" is true only because if I hadn't hit the ball, it wouldn't have moved.

Generalising from this example, he infers:

- (53) "*C* caused *E*" is true only because *E* counterfactually depends on *C*.

As we have discussed, *only because* implies counterfactual dependence. And indeed, (53) clearly implies that if *E* didn't counterfactually depend on *C*, "*C* caused *E*" would be false. Thus Hume comes to believe that counterfactual dependence is necessary for causation. It is an honest mistake, the result of perfectly rational principles of linguistic behaviour.

Moreover, if Hume takes sufficiency for granted, he will also come to believe that counterfactual dependence is necessary and sufficient for causation. And so he writes in his *Enquiry* the following sentence: "we may define a cause to be *an object followed by another, ... where, if the first object had not been, the second had not existed.*" As we have seen, our proposal accounts for the cases where this idea goes wrong. But more than that, we account for why we thought it was right.

Chapter 3

Imaginative structures

Human rationality depends on imagination. People have the capacity to be rational at least in principle because they can imagine alternatives.... The principles that guide the possibilities people think of are principles that underpin their rationality.

— Ruth Byrne, *The Rational Imagination* (2005:29)

3.1 Introduction

A truly remarkable achievement of reasoning is the ability to consider hypothetical scenarios, and to have others imagine the same scenario that one has in mind.¹ This ability to create, and coordinate, on hypothetical scenarios plays an essential role in the interpretation of conditionals and causal claims. For it is often thought that the truth conditions of these constructions depends not only on what actually happens, but on what happens in certain hypothetical scenarios as well.

Here is an illustration (adapted from an example by Schulz 2007:101). Switches A and B are connected to a light. Each switch can be either up or down. As the wiring of Figure 3.1 depicts, the light is on just in case both switches are up. Currently, switch A is down, switch B is up, and the light is off.

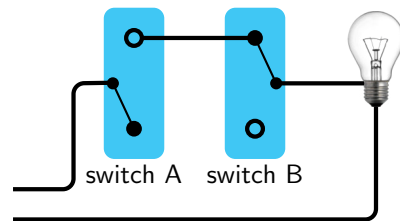


Figure 3.1

Consider (1) in this context.

- (1) a. The light is off because switch A is down.
b. If switch A were up, the light would be on.

The sentences in (1) are straightforwardly true. In theory, one could try to reject them, arguing that “If switch A were up, the light would still be off, since switch

¹A much shorter version of this chapter has been published as McHugh (2022).

B would be down.” In practice, however, no one responds in this way. We somehow manage to coordinate on what hypothetical scenarios to consider when we interpret these sentences. We manage to do this without explicit training; unlike how children learn, say, arithmetic, no child has to be explicitly taught what hypothetical scenarios to imagine in response to a sentence. And we often manage to coordinate on what hypothetical scenarios to consider in situations we have never encountered before. This is an ordinary, extraordinary fact about reasoning, one we would like to understand.

In this chapter we provide a general framework for hypothetical reasoning, with the specific goal of predicting what scenarios we consider when we interpret a conditional or causal claim. Our data are judgements about the acceptability of these sentences. The fundamental idea of our approach is that when we interpret a conditional or causal claim, we identify a part of the world to change, and imagine changing that. Sentences – such as conditional antecedents and *because* clauses – are about parts of the world, parts we allow to change when we interpret them. To determine whether a conditional or causal claim is true we look to the possible futures after this change.

In section 3.2 we compare the semantics of these conditionals and causal claims, proposing that they both make use of the same general faculty of hypothetical reasoning. Section 3.3 formalises the operation of imagining a part of the world changed. We present some advantages of this approach in section 3.4 and showcase its expressive power in section 3.5. Section 3.6 introduces a relation between language and the world that tells us what part of the world to change in response to a sentence, and sections 3.7 and 3.8 explore what inference patterns our proposal validates. In Chapter 4 we consider how to generalise the framework to logically complex sentences and explore their consequences for the logic of conditionals. Section 4.4 shows the advantages of our approach over two previous frameworks, those using a similarity order over worlds and premise semantics.

3.2 Comparing conditionals and causal claims

A compelling idea is that when we interpret conditionals and causal claims we make use of the same general ability to reason hypothetically. In other words, sentences such as conditional antecedents or *because*-clauses raise hypothetical scenarios in a uniform way. One piece of evidence for this is that we so often paraphrase one kind of sentence with the other. An early example is from David Hume, who inaugurated so-called ‘counterfactual approaches to causation’ in 1748 when he paraphrased a causal claim using a conditional:

We may define a cause to be an object followed by another ... where,
if the first object had not been, the second never had existed.

An Enquiry concerning Human Understanding, Section VII

Ramsey (1929a:17) went so far as to write that “*because* is merely a variant on *if*, when p [the conditional antecedent] is known to be true” – an idea that pushes the parallel between conditionals and causality past its breaking point. Nonetheless, the correspondence lives on today in the panoply of counterfactual approaches to causation (see e.g. Collins, Hall, and Paul 2004, Menzies and Beebe 2020).

Here is a more contemporary example of the close connection between conditionals and causal claims. Title VII of the US 1964 Civil Rights Act states,

It shall be an unlawful employment practice for an employer to fail or refuse to hire or to discharge any individual ... because of such individual’s race, color, religion, sex, or national origin.

(78 Statute 241, Sec. 703(a)(1), p. 255)

The text uses a causal word: *because*. Now here is Justice Elena Kagan in 2019, discussing the same law during oral argument for a US Supreme Court case, *Bostock v Clayton County*.

Kagan: What you do when you look to see whether there is [sex] discrimination under Title VII is, you say, would the same thing have happened to you if you were of a different sex?

(Oral argument, pp. 41–42)

Notice how Kagan uses a conditional (“if you were of a different sex”) to express a causal claim (“because of ... sex”). As a matter of fact, Title VII does not contain any conditionals of the kind uttered by Kagan. A causal claim winds up expressed as a conditional. Paraphrasing one as the other happens all the time. For another example, in 2020 Sauntore Thomas went to the bank to deposit a perfectly valid cheque. The bank refused. Instead they called the police and launched a fraud investigation. Asked about his experience in the *Detroit Free Press*, Thomas replied, “They discriminated against me because I’m black. None of this would have happened if I were white”. Thomas moves seamlessly between a causal claim and a conditional. The causal claim and the conditional appear to be saying the same – or almost the same – thing.

This suggests that the meaning of causal claims and conditionals have a great deal in common. If conditionals and causal claims did not share a meaning component, the fact that we so often use one to argue for the truth of the other – as David Hume, Elena Kagan and Sauntore Thomas did – would be a mystery.

We propose that the interpretation of conditionals and causal claims raises hypothetical scenarios in a uniform way, making use of the same general capacity to imagine a situation changed so that a given sentence, such as a conditional antecedent or cause argument of a causal claim, is true in it.

The uniformity of hypothetical reasoning has the added bonus of considerably simplifying research into conditionals and causal claims. For we do not have to give two separate accounts of how we construct hypothetical scenarios, one for conditionals and one for causal claims. A single account will suffice.

If such a uniform programme is to succeed, there is one main obstacle to overcome: the status of sufficiency, which we turn to now.

3.2.1 Sufficiency and hypothetical reasoning

In Chapter 2 we saw evidence that *cause* and *because* imply that the cause was in some sense sufficient for the effect. For example, recall the robot case from section 2.2, where the robot has to get to Main Street, and turns at random at each fork. Actually, it took First Street and then Road B. Consider (2).

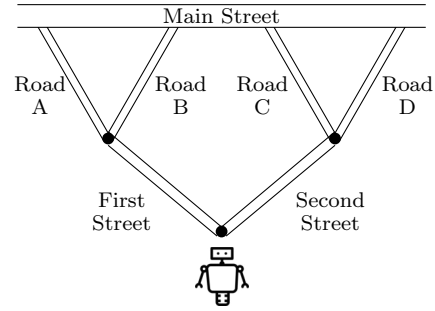


Figure 3.2

- (2)
 - a. The robot took Road B because it took First Street.
 - b. The robot taking First Street caused it to take Road B.

(2) are intuitively unacceptable in this context.

Suppose we minimally change the scenario to make taking Road B sufficient for taking First Street, say, by programming the robot to always change direction (e.g. if it turns left at one fork it must turn right at the next). In this context, suddenly (2) are acceptable. This suggests that when the robot turned at random, (2) were unacceptable due to a failure of sufficiency.

We can account for this by incorporating the openness of the future in the analysis of sufficiency.² When we evaluate (2), we hold fixed the facts that occur prior to the robot taking First Street and allow facts after that to vary. Since the robot took Road B after taking First Street, we do not fix the fact that the robot took Road B. Then as it was possible for the robot to take Road A instead of Road B after First Street, we predict that taking First Street was not sufficient for the robot to take Road B. Given that *cause* and *because* entail that the cause was sufficient for the effect, we predict (2) to be unacceptable.

However, some violations of sufficiency do not involve the openness of the future. Suppose Ali was born in Ireland and has an Irish passport, and recall (3).

- (3)
 - a. Ali has an Irish passport because he was born in Europe.
 - b. The fact that Ali was born in Europe caused him to get an Irish passport.

There is something intuitively wrong with these sentences. They greatly improve

²Previous proposals that use the openness of the future in the semantics of conditionals include Thomason (1970), Thomason and Gupta (1980), Condoravdi (2002), Arregui (2005), Ippolito (2013), Khoo (2015), and Canavotto (2020).

when we replace the cause with a minimally different one that is sufficient for the effect.

- (4) a. Ali has an Irish passport because he was born in Ireland.
- b. The fact that Ali was born in Ireland caused him to get an Irish passport.

This suggests that causal claims imply that the cause is in some sense sufficient for the effect. The question is how to analyse this notion of sufficiency. It cannot be entailment, where A entails C just in case in every logically possible world where A is true, C is true. Take, for example, (5).

- (5) a. The light turned on because I flicked the switch.
- b. Flicking the switch caused the light to turn on.

We can accept these even though flicking the switch does not entail that the light turns on. In a world where, say, the electricity is down, we can flick the switch without the light turning on. When we evaluate (5) we fix the fact that the electricity is working.

A natural alternative is that A is sufficient for C just in case the conditional “if A , would C ” is true. One problem with this idea is that the most prominent theories of the meaning of *would*-conditionals (e.g. Stalnaker 1968, Lewis 1973b, Kratzer 2012) predict that if sentences A and C are both true, then the conditional “if A , would C ” is true too.³ *Would*-conditionals, in other words, are ‘strongly centered’ around the world of evaluation: when A is actually true, the only hypothetical scenario A raises is the actual one. For similarity-based approaches to conditionals (e.g. Stalnaker 1968, Lewis 1973b, Pollock 1976) this is a natural constraint to impose – surely every world is more similar to itself than any other world is to it. So if A is actually true, the most similar world to the actual world where A is true is just the actual world itself. A similar kind of ‘strong centering’ is part of Kratzer’s approach to conditionals, as we will see in section 4.4.1.

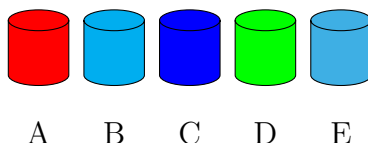
The problem is that, as (3) show, when we interpret a causal claim we consider various ways for the cause to hold, even when we know in what particular way the cause in fact holds. For example, there are many ways to be born in Europe (being born in Ireland, Hungary, Ukraine, and so on). Intuitively we consider all of these ways when interpreting (3), even though we are told that Ali was in fact born in Ireland.

To ensure that the phenomenon we are studying is robust, let us consider more examples. Compare the (a)- and (b)-sentences below.

- (6) *Alice is 20 years old. The legal drinking age is 18.*

³See also Mandelkern (2018), Cariani and Santorio (2018), and Cariani (2021) for arguments that *will*-conditionals are strongly centered.

- a. Alice can order wine because she's over 12.
 - b. Alice can order wine because she's over 18.
- (7) *Yves visits an art supply shop looking for a can of ultramarine paint. Ultramarine is a particular shade of blue, and the only colour Yves paints with. He sees the paint cans below. Paints B, C and E contain different shades of blue. He buys paint C, which contains ultramarine paint.*
- a. Yves bought paint C because it is blue.
 - b. Yves bought paint C because it is ultramarine.



- (8) *Bob has 1000 euro. Sarah has 2000 euro.*
- a. Sarah is richer than Bob because she has more than 500 euro.
 - b. Sarah is richer than Bob because she has more than 1000 euro.
- (9) *Let x and y be numbers, where $x \neq 0$ and $y = 0$.*
- a. xy is 0 because y is less than 10.
 - b. xy is 0 because y is 0.

The (a)-sentences are much worse than the (b)-sentences. In each case we would naturally say that the cause in the (a)-sentence is not sufficient for the effect, while the cause in the (b)-sentence is.

We cannot account for these contrasts using the openness of the future as we did in the robot case. In (6)–(9), the causes in the (a)- and (b)-sentences occur at the same time. For example, in (7) presumably the paint became blue at the same time it became ultramarine. If we tried to account for the unacceptability of (7a) using branching futures, we would have to say that there was a time when the paint was blue but not yet determined that it would be ultramarine. But paint makers do not make paints by adding the ‘general’ pigments first and then the ‘specific’ pigments. ‘General’ pigments do not even exist – every blue is a kind of blue.

We also cannot account for these contrasts using counterfactual dependence. In both the (a)- and (b)-sentences, the effect counterfactually depends on the cause. For example, if Alice were under 18 she could not order wine, and if she were under 12 she certainly could not order wine.

Rather, (6)–(9) suggest that when we interpret *because*, we consider various ways for the cause to be true. There are many ways to be over 12 years old (13, 14, 15, ...), many kinds of blue, many ways for a number to be less than 10 (9, 8, 7, ...), and so on. The unacceptability of the (a)-sentences tells us that *cause*

and *because* require that the effect hold under every way for the cause to hold.

The question is then how to understand what are the various ‘ways’ for a sentence to be true. As we have seen, we cannot say that *A* is sufficient for *C* just in case the conditional “if *A* would *C*” is true, since *will/would*-conditionals are strongly centered but sufficiency is not. We will not review the arguments for strong centering in conditionals here, but we will consider two environments where *will/would* conditionals exhibit a different modal force than *cause* and *because*. We will see that *cause* and *because* instead pattern with conditionals using *guaranteed* in place of *will/would*. The environments we consider are betting contexts and probability statements.

3.2.2 The modal force of *will/would* and *cause/because*

Bets. Consider the robot context, where the robot chooses at random between First and Second Street. If it takes First Street it chooses at random between Roads A and B, if it takes Second Street it chooses at random between Roads C and D. The robot has not set out on its journey yet. Four friends come along who do not know how the robot decides which way to go (e.g. whether it chooses at random or its route is pre-programmed). They make the following bets.⁴

- (10) a. Alice: “I bet that if the robot takes First Street, it will take Road B because it took First Street.”
- b. Bob: “I bet that if the robot takes First Street, taking First Street will cause it to take Road B.”
- c. Chandi: “I bet that if the robot takes First Street, it will take Road B.”
- d. Darius: “I bet that if the robot takes First Street, it is guaranteed to take Road B.”

On this particular occasion, the robot took First Street and then Road B. The four friends are then told that the robot decided which way to go at random, and in particular, that after taking First Street the robot could have taken Road A rather than Road B. Now, who won their bet?

Intuitively, it is clear that Alice, Bob and Darius lost their bets. I would also say that Chandi won her bet. Suppose the others refuse to give Chandi her winnings on the basis that the robot could have taken Road A instead of Road B. As Prior (1976) pointed out, Chandi can sensibly reply, “I didn’t bet that if the robot takes First Street, it *must* take Road B. Rather, I bet that if it takes First Street, it *will* take Road B. And it did take Road B. So I was right.” I am persuaded by Chandi’s argument that she won her bet.

⁴Examples featuring *will/would* conditionals in betting contexts have been previously discussed by Prior (1976:100), Moss (2013), Belnap, Perloff, and Xu (2001:160), Cariani and Santorio (2018) and Cariani (2021:63).

We can check whether the bet using *guaranteed* is false for the same reason as the bets using *because* and *cause*; namely, due to a failure of sufficiency. As we did when discussing sufficiency above, imagine instead that the robot was programmed to always change direction. In that context, intuitively Alice, Bob and Darius also won their bets.

The shows that the sufficiency inference from *because* and *cause* can be paraphrased with a conditional featuring *guaranteed* but not by one featuring *will*.

We see the same pattern with *would*-conditionals. Given that the robot turned at random (and actually took First Street), consider the following statements.

- (11)
- a. Idris: “I bet that if the robot had taken Second Street, it would have taken Road C because it took Second Street.”
 - b. Javier: “I bet that if the robot had taken Second Street, it taking Second Street would have caused it to take Road C.”
 - c. Khalil: “I bet that if the robot had taken Second Street, it would have taken Road C.”
 - d. Lina: “I bet that if the robot had taken Second Street, it would have been guaranteed to take Road C.”

Who won their bet? Intuitively, the outcome of Khalil’s bet remains undecided, while Idris, Javier and Lina lost. Suppose we can open up the robot to inspect how it makes decisions. We find the part showing that it turns at random (say, a Geiger counter checking whether a radioactive substance has decayed) and present it to those who bet. This does not resolve whether Khalil won or lost, but the fact that the robot turned at random is enough for the others to lose.

Probability statements. Consider again the robot context, where the robot turns at random. The robot will soon begin its journey. Our friends from above make the following statements.

- (12)
- a. Alice: “If the robot takes First Street, it will take Road B because it took First Street.”
 - b. Bob: “If the robot takes First Street, it taking First Street will cause it to take Road B.”
 - c. Chandi: “If the robot takes First Street, it will take Road B.”
 - d. Darius: “If the robot takes First Street, it is guaranteed to take Road B.”

For each of these people, what is the probability that what they said is true?⁵

⁵This way of formulating the question follows a strategy by Mandelkern (2018), to avoid the objection that conditionals cannot be assigned probabilities. For when one is asked “What is the probability that what *x* said is true?” where *x* said a conditional, it is clear one is asked to assign a probability to an entire conditional; rather than, say, assign a probability to the consequent under the assumption of the antecedent. For discussion see Mandelkern (2018:304–

Intuitively, what Chandi said has a 50% probability of being true, but for each of the others, what they said has a 0% probability of being true.

Similar observations hold for past tense conditionals. Consider the following, given that the robot decides which way to turn at random and took First Street.

- (13) a. Idris: “If the robot had taken Second Street, it would have taken Road C because it took Second Street.”
 b. Javier: “If the robot had taken Second Street, it taking Second Street would have caused it to take Road C.”
 c. Khalil: “If the robot had taken Second Street, it would have taken Road C.”
 d. Lina: “If the robot had taken Second Street, it would have been guaranteed to take Road C.”

For each of the speakers above, what is the probability that they they said is true? Similar to above, the intuitive response here is that there is a 50% probability that what Khalil said is true, and for each of the others, a 0% probability that what they said is true.

Thus in judgements of bets and probability, *cause* and *because* pattern with conditionals using *guaranteed* and not with conditionals using *will* or *would*. This is evidence that *cause* and *because* quantify over a wider range of hypothetical scenarios than *will/would* conditionals do: *will/would*-conditionals involve strongly centered modality while *cause* and *because* do not. Our hope for a uniform account of how conditionals and causal claims raise hypothetical scenarios looks doomed.

3.2.3 A way out: selection functions

There is, however, a way out. For there is no evidence that the strong centering requirement of conditionals must come from our general capacity to imagine hypothetical scenarios. Indeed, Cariani and Santorio (2018), Cariani (2021), and Mandelkern (2018) propose that strong centering has another source: there is a separate component in the meaning of *will/would* conditionals, called a selection function, that chooses a world from a set of worlds. The only additional constraint on the selection function is bias toward actuality: if the selection function can choose the world of evaluation, it must.

Selection functions allow us to rescue a uniform account. For we can propose that we have a general capacity to consider hypothetical scenarios in response to a sentence – one we use to interpret both conditionals and causal claims – that is not strongly centered. In addition, the interpretation of *will* and *would*, but not causal claims, involves a selection function, accounting for strong centering.

- (14) A selection function $s : (W \times \wp(W)) \rightarrow W$ takes a world w and a set of worlds p and returns a world, where
- a. *Success*: $s(w, p) \in p$
 - b. *Centering*: If $w \in p$ then $s(w, p) = w$.⁶

We will make the following assumption about the relationship between the selection function and how people respond to utterances of conditionals.

- (15) **Proposal.** A *will/would*-conditional is assertable only if it is true on every selection function.

In other words, conversational participants assume that the truth of a *will/would*-conditional does not depend on the selection function. To illustrate using the classic example from Quine (1950), each selection function determines which of the following pair is true and which is false (assuming Bizet and Verdi could not both have come from some third country).

- (16) a. If Bizet and Verdi had been compatriots, Bizet would have been Italian.
 b. If Bizet and Verdi had been compatriots, Verdi would have been French.

But both are unassertable since their truth depends on the selection function.

I presume that we adhere to the assertability condition in (15) because the selection function is an unknowable parameter of interpretation; apart from the success and centering constraints, it picks a world at random.

It is natural to wonder why there is a selection function in the meaning of *will* and *would* at all given that we cannot identify it. Following ideas by Cariani (2021), I propose that the selection function is part of the semantics of *will* and *would* because it ensures that we can talk about – e.g. make bets and conjectures about, ask questions about, weigh up probabilities regarding – the actual future. We need some way to talk about the actual future, even if it is currently unknowable which future will in fact come to be actual. After all, the actual future is the only (epistemically/nomically/logically...) possible future we get to experience. English chooses to give this important communicative role to *will/would*. The selection function's centering condition fulfils our need to talk about the actual future, but when it is unknowable which world will come to be actual, it is unknowable which selection function is in use.

⁶There are various ways to model what happens when p is empty, i.e. the inconsistent proposition. Stalnaker (1968) assumes that the selection function returns the absurd world, where every proposition is true. Cariani and Santorio (2018) require that $s(w, p) \in p$ if p is nonempty. Another option is to assume that when we interpret *will/would* we presuppose that s is a selection function. When p is empty the success condition is violated, so s cannot be not a selection function, resulting in presupposition failure.

Turning to *would*, following Abusch (1997, 1998) we assume that *will* and *would* decompose into a modal element – what Abusch calls WOLL – plus a tense morpheme: *will* is WOLL + PRESENT; *would* is WOLL + PAST. Standardly, the semantics of tense does not deal with selection functions, so we assume the selection function is contributed by WOLL and is therefore present in the semantics of both *will* and *would*. We also assume, following standard semantics of tense, that tense does not alter modal force – that is not in tense’s job description, so to speak – so if *will* picks out a single world, *would* must too. Considerations of compositionality favour an account where WOLL makes the same semantic contribution wherever it occurs. For whatever reason, many languages including English find it economical to use the same morpheme to talk about the actual future and counterfactual possibilities. In languages that do so, the strangeness of conditional excluded middle in counterfactuals is the price we pay for being able to talk about the actual future.

3.2.4 Adding time

We will determine the set of hypothetical scenarios raised by *cause* and *because* in two steps. Formally, we begin with a set of *moments*, which represent snapshots of the world at a moment in time. A *world* is defined as a linearly ordered set of moments. The set of logically possible worlds is the set of linear orders of moments. From this set we designate a subset of nomically possible worlds P .

We propose that changes are calculated at a moment in time, so there is a function $int(t, A)$ – what we may call an intervention function – that takes a moment t and a sentence A and returns a set of moments. We then ‘play the laws’ from those moments according to the laws of nature. Interventions leave the past of t untouched, but can change the present and future. The presents that result from the intervention at a moment are $int(t, A)$, while the futures that result from the intervention are the futures of the presents that are nomically possible.

Let us make this discussion precise. Where M is a set of moments, a world is a pair (M, \preceq) where $\preceq \subseteq S \times S$ is a linear order, i.e. it is transitive, antisymmetric, and connex: $t \preceq t'$ or $t' \preceq t$ for all $t, t' \in M$. \preceq represents time. For any world w and moment t on w , let $w_{\prec t}$ be the segment of w up to but not including t , let $w_{\succeq t}$ be the sequence of w from t on, including t , and let \frown denote concatenation.⁷ And where $w = (S, \preceq)$, let us write $t \in w$ to mean $t \in S$.

Then given a moment t , the set of possible futures of t is $\{w_{\succeq t} : w \in P, t \in w\}$. To construct the modal horizon of sentence A at a moment t , we take the possible

⁷Formally, we define concatenation as follows. Where $= (X, \preceq)$ and (Y, \preceq') are linear orders, $(X, \preceq) \frown (Y, \preceq') = (X \uplus Y, \preceq \frown \preceq')$ where $X \uplus Y$ is the disjoint union of X and Y , and $x \preceq \frown \preceq' y$ just in case (i) $x, y \in X$ and $x \preceq y$, or (ii) $x, y \in Y$ and $x \preceq' y$, or (iii) $x \in X$ and $y \in Y$.

futures of the moments in $\text{int}(t, A)$ and glue on the actual past.⁸

(17) For any world w , sentence A , set of worlds P and moment t , define

$$mh_{P,t}(w, A) := \{w_{\prec t} \frown w'_{\succeq t'} : w' \in P, t' \in \text{int}(t, A), t' \in w'\}.$$

For any sentence S , let $|S|$ be the set of worlds where S is true. Let mh be a function from worlds and sentences to sets of worlds, and s be a selection function. We propose the following clauses for sufficiency (\gg) and *will/would*-conditionals ($>$), for any world w and sentences A and C .⁹

⁸By gluing on the actual past, our account follows approaches to modality using a historical modal base (Thomason 1970, 2014, Condoravdi 2002, Kaufmann 2005, Ippolito 2013, Cariani and Santorio 2018, Cariani 2021). In addition to the arguments provided by these authors for a historical modal base, a further argument for gluing on the actual past is that can intervene a moment and yet evaluate sentences about things before that moment.

- (i) If a rabbit had suddenly appeared in this hat, Caesar would still have crossed the Rubicon.

It seems we can interpret (i) by intervening recently, and then looking back to the state of the world prior to the intervention time. Here is a naturally occurring example (Elizabeth Evangelina, [Quora](#)).

- (ii) The reason Tom is a psychopath, is because he was born of a love potion. However, even if his dad stayed, he would have still been a psychopath because he would still have been born of a love potion.

If we left out the actual past, there is a worry that we would lose the ability to evaluate facts that occurred prior to intervention time. We can evaluate such sentences, despite their strangeness. (I presume they are strange because they violate Condoravdi's (2002:83) diversity condition. Here the antecedent is redundant: if the truth value of the sentence only depends on the actual past, why raise hypothetical scenarios needlessly? Why roam logical space when you can already find what you are looking for at home?)

⁹The clauses in (18) are, of course, a simplification of the compositional semantics of conditionals. Let us rewrite the clauses in terms of a more plausible account: von Fintel's (1994:87–89) formalism of the restrictor view, coupled with the notion of a modal horizon from von Fintel (2001b). On this account, *if*-clauses restrict modals via coindexing of a domain variable. Each domain variable is assigned a set of worlds. Written in terms of that account, the clauses in (18) are as follows, where P is the set of nomic possibilities and t the intervention time, and we let $\llbracket \alpha \rrbracket^{g,mh,s} = \{w' : \llbracket \alpha \rrbracket^{w',g,mh,s} = 1\}$.

- (i) a. $\llbracket \beta \gg \alpha \rrbracket^{w,g,mh,s}$ is defined only if $mh_{P,t}(w, A) \subseteq mh$.
 b. If defined, $\llbracket \beta \gg \alpha \rrbracket^{w,g,mh,s} = 1$ iff $mh \cap \llbracket \beta \rrbracket^{g,mh,s} \subseteq \llbracket \alpha \rrbracket^{g,mh,s}$.
- (ii) $\llbracket \text{if}_c \beta, \alpha \rrbracket^{w,g,mh,s} = \llbracket \alpha \rrbracket^{w,g',mh',s}$, where
 a. $g'(c) = g(c) \cap \llbracket \beta \rrbracket^{g,mh,s}$ and $g(x) = g'(x)$ for all variables $x \neq c$;
 b. $mh' = mh \cup mh_{P,t}(w, \beta)$.
- (iii) a. $\llbracket \text{will}_c \alpha \rrbracket^{w,g,mh,s}$ is defined only if s is a selection function.
 b. If defined, $\llbracket \text{will}_c \alpha \rrbracket^{w,g,mh,s} = 1$ iff $s(w, mh \cap g(c)) \in \llbracket \alpha \rrbracket^{g,mh,s}$.

- (18) Where P is the set of nomically possible worlds, t the intervention time, and s the selection function,

$$\begin{aligned} A \gg C \text{ is true at } w & \quad \text{iff} \quad mh_{P,t}(w, A) \cap |A| \subseteq |C| \\ A > C \text{ is true at } w & \quad \text{iff} \quad s(w, mh_{P,t}(w, A) \cap |A|) \in |C| \end{aligned}$$

3.3 A change of world

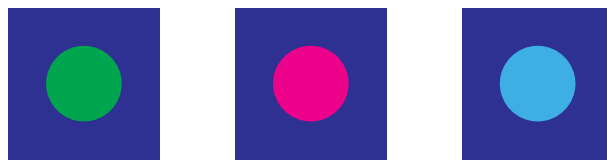
When we are asked to imagine the world changed so that a given sentence is true, intuitively we identify the part of the world that needs to change, and change that. In the light switch example above, when we are asked to imagine switch A up, we imagine changing the position of switch A. Everything else is background – the *ceteris* in *ceteris parabus* (the ‘other things’ in ‘other things being equal’). Intuitively we do not rank all possible worlds according to an order, and we do compare all possible worlds in terms of what propositions they make true. Rather, we imagine a part of the world changed.

How do we know what part to change? To answer this question, consider the following image:



Imagine the circle with a different colour. What images come to mind?

Some images we may imagine are:



Some images we do not imagine are:



Remarkably, we all seem to imagine the same kinds of images. Notice that we are only given the image and the sentence “imagine the circle with a different colour”. We are not explicitly told what stays the same: we figure that out

ourselves. And we all tend to figure it out in the same way. We would like to explain this systematic behaviour of our imagination.

One might try to explain it by saying that we all share the same concept of similarity, ranking all possible images by how similar they are to the actual image, and imagine the most similar images to the image we are given. Or one might say that we all identify the same set of relevant propositions that the image makes true, and seek to preserve the truth of as many of these propositions as possible that while maintaining consistency with the circle being a different colour.

These ideas, while promising, leave most of our question unanswered. We would still have to explain why we all seem to have the same similarity order, or select the same set of relevant propositions. One way to answer this would be to identify what determines the similarity order or what propositions count as relevant. For if we had an answer to this question, we could explain that we all tend to imagine the same images because the similarity order, or set of relevant propositions, is determined by the scenario itself and the sentence we are asked to imagine true – information that is available to everyone who is shown the image and asked to imagine the circle a different colour.

But on second thought, when we are asked to imagine the circle a different colour, intuitively we do not rank all images by similarity, and we do not compare all images in terms of what propositions they make true. Rather, we identify a part of the image that needs to change, and change that. Let us therefore start with a simple question: when we imagine the circle a different colour, what parts of the image change, and what parts stay the same?

We can answer this question without much thought at all. Let us begin by listing parts of the image, and checking one by one whether each part changes when we imagine the circle a different colour. This procedure is illustrated in Figure 3.3. When we sort these parts according to whether they change or stay the same, we end up with something like Figure 3.4. What property do the parts that change have in common? Looking at Figure 3.4, we see that they all overlap the circle. And the parts that do not overlap the circle stay the same.

- (19) For any part of the image x , if x does not overlap the circle, x stays the same.

This gives us the *ceteris* in *ceteris paribus* – the ‘other things’ in ‘other things being equal’. In this case, the *ceteris* are the parts of the image that do not overlap the circle.

But what does it mean for a part of the image to ‘stay the same’ when we imagine the circle a different colour? Intuitively, it is for that part to also be part of what we imagine.

- (20) For any part of the image x , if x stays the same when we imagine the circle a different colour, x is part of every image we imagine when we imagine the circle a different colour.

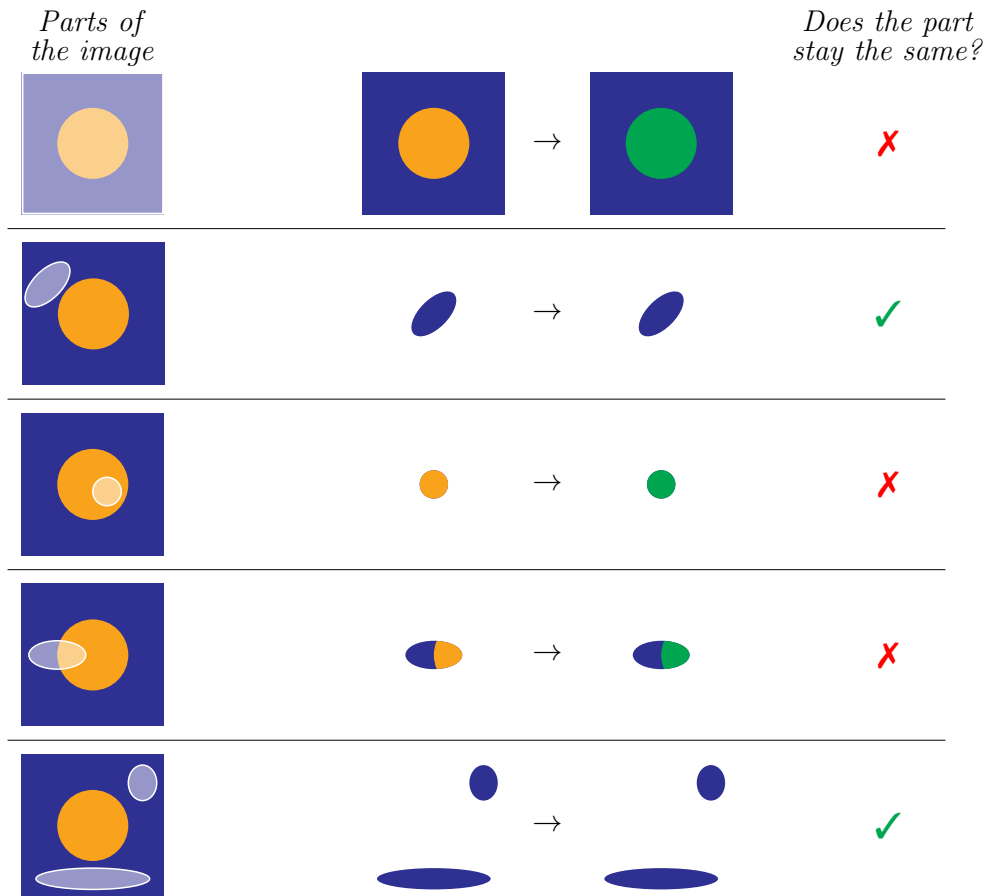
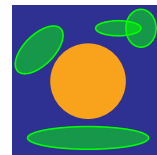


Figure 3.3



Some parts that change.



Some parts that stay the same.

Figure 3.4

This gives us the *paribus* in *ceteris paribus* – the ‘being equal’ in ‘other things being equal’. *Paribus* means having the *ceteris* as part.

Putting (19) and (20) together, we have:

- (21) For any part of the image x , if x does not overlap the circle, x is part of every image we imagine when we imagine the circle a different colour.

We are asked to imagine the circle a different colour. Given the images satisfying

(21), let us restrict to those images where the circle is a different colour.

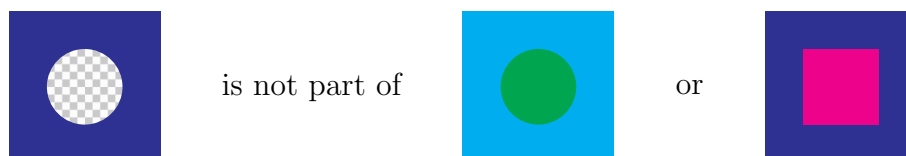
(22) When we imagine the circle a different colour, in every image we imagine, the circle is a different colour.

These two principles, (21) and (22), give us the results we are looking for. The images that we imagine when we imagine the circle a different colour are all and only those satisfying (21) and (22). This is stated in (23).

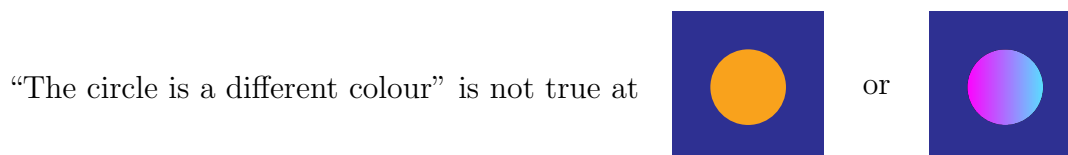
(23) Where i is the image we started with – the orange circle on a blue background – when we are asked to imagine that the circle has a different colour, an image i' is among the images we imagine if and only if

- every part of i that does not overlap the circle is part of i' , and
- “the circle has a different colour” is true at i' .

(21) rules out the first two images we do not imagine – the images where the background is a different colour, and where the circle becomes a square – since there are parts of the original image not overlapping the circle that fail to be part of them; respectively, the whole dark blue background in the first image, the corners around the circle in the second image (where the grey check pattern represents the absence of a part).



(22) rules out the second two images we do not imagine – the original image and the image with many colours. The first is ruled out since the circle is not a different colour. The second is ruled out because the circle has many colours, given that we interpret “a different colour” as “exactly one different colour”.

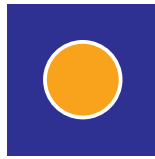


Given (23), we can account for why, when asked to imagine that the circle has a different colour, we all tend to imagine the same kind of images. For (23) is formulated using properties we all agree on. Everyone agrees on what parts of the image do not overlap the circle, and “the circle has a different colour” means the same thing to different people.

3.3.1 Coarser representations of the image

In the above discussion, we considered all parts of the image whatsoever. One may worry about the generality of our results, since that is not the only way to represent the image. We may, for example, consider coarser representations. In this subsection I wish to illustrate just how general the rule in (21) is, by showing that it makes the right predictions for various representations of the image, while alternative rules do not.

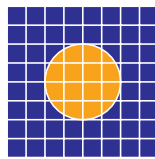
One way to represent the image, for example, is to think of it as made up of only two parts, the circle and the background:



If these are the only two parts of the image we consider, we can write the rule as:

(24) Rule 1. When changing x , if y is not identical to x , y stays the same.

Now, there is no reason to suppose that we cannot take a more fine-grained perspective; taking in account, for example, the parts of the circle. A more general approach is to partition the image, as a mosaic artist does with tiles, or as a printer or computer screen does with pixels, or as researcher does with random variables, where each variable corresponds to a region of the image, their values corresponding to colours. Then the above representation – dividing the circle in two – falls out as a special case; namely, a partition whose only members are the circle and the background. Partitions work provided the partition is fine-grained enough to recover the image (i.e. no cell of the partition overlaps both the circle and the background); for example:



Rule 1 gives the wrong results here.¹⁰ Consider any tile that is part of the circle. Since it is not identical to the whole circle, Rule 1 wrongly predicts that it stays the same when we imagine the circle a different colour. Let us update our rule to work for partitions. Our Rule 2 is the following, where y ranges over cells of the partition.

¹⁰The are some striking parallels between the present observations and a previous discussion concerning the meaning of *only* (see e.g. Heim 1990, von Fintel 1997:10–14 and Fox 2003). I leave this as a suggestive remark for now.

- (25) Rule 2. When changing x , if y is not part of x , y stays the same.

Of course, in addition to considering the cells of the partition, we can also consider the parts of the image that are made up of these cells, such as the entire image. Rule 2 does not work in this case: since the entire image is not part of the circle, according to Rule 2 it stays the same when we change the circle, which is incorrect. A rule that also works for combinations of cells is the following (where y is not comparable with x just in case x is not part of y and y is not part of x).

- (26) Rule 3. When changing x , if y is not comparable with x , y stays the same.

Again, we are also capable of taking an even more fine-grained perspective, considering, say, the parts of the image that overlap the circle and background. For instance, the top half of the image is not part of the circle, nor the circle part of it, so Rule 3 wrongly predicts that the top half stays the same when we imagine the circle a different colour.

As we saw in (21), when we consider all parts of the image whatsoever, the correct rule is:

- (27) Rule 4. When changing x , if y does not overlap x , y stays the same.

Each rule is weaker than the one before: Rule 1 implies Rule 2, which implies Rule 3, which implies Rule 4 (since identity implies parthood, which implies comparability, which implies overlap). The point I wish to make is that Rule 4 works in all the cases we have just considered. Even though the motivation for Rule 4 came from looking at all parts of the image whatsoever, it also gives the right results for coarser representations. Rules 1, 2 and 3 do not enjoy such generality, only giving the right results when we restrict the kinds of parts we consider. More precisely, the rules agree under the following assumptions (where to be atomic is to have no proper parts).

- (28) a. If x is atomic, Rules 1 and 2 are equivalent.
 b. If everything that overlaps x is part of x , Rules 2 and 3 are equivalent.
 c. If everything that overlaps x is comparable with x , Rules 3 and 4 are equivalent.

The level of generality offered by Rule 4 is crucial if we are to find a rule that captures how we imagine changes not only in simple cases – such as changing the colour of a circle – but when we are asked to imagine changing a part of any situation whatsoever.

3.3.2 Proposal: varying a state

In full generality, then, let us assume that we have a set S of entities, called states, and a parthood relation \leq between them. Following Fine (2017b), the pair (S, \leq) we call a *state space*. At a minimum, we assume that parthood is reflexive, transitive and anti-symmetric. We will also assume that every state is part of a moment, where a moment is defined as a state that is not a proper part of any state (in the terminology of Fine (2017b), our state spaces are *world spaces*). As usual, we define that two states overlap (denoted $s \circ t$) just in case they have a part in common and are disjoint (denoted $s \not\circ t$) just in case they do not.¹¹

We are interested in what states we imagine when we vary a particular state. To that end, let us introduce the following terminology.

- (29) **Definition** (*s*-variant). For any moments t, t' and state s , define that t' is an *s*-variant of t just in case every part of t that does not overlap s is part of t' .

$$t' \text{ is an } s\text{-variant of } t \quad \text{iff} \quad \forall u \leq t (s \not\circ u \Rightarrow u \leq t')$$

We propose the following; in effect, that Rule 4 is the only constraint governing what states we imagine when we vary a state.

- (30) **Proposal.** For any moments t, t' and state s , when we imagine changing s at t , t' is among the moments we imagine iff t' is an *s*-variant of t .

It turns out that one can greatly simplify the definition of *s*-variant by imposing some natural constraints on parthood. In the simplified definition, one need not consider *all* parts of a moment that do not overlap s , but instead only a single part of it. In the next section we give this simplification. Then in section 3.4 we will put our proposal in (30) to the test by considering a range of examples.

3.3.3 Simplifying the definition of what stays the same

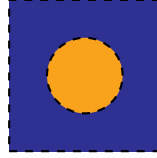
Consider again the image of the circle in a square, where we are asked to imagine the circle a different colour. We have identified what *parts* stay the same: those that do not overlap the circle. But now suppose we are asked to find *the* part of the image that stays the same. What part should we pick?

Even though many parts stay the same, we still have an answer this question. Intuitively, the part that stays the same is the *largest* part that stays the same.

¹¹Note we could work with overlap directly, rather than defining it from parthood. If parthood is extensional (no distinct states with proper parts have all the same proper parts), one may recover parthood from overlap by defining that x is part of y just in case every state that overlaps x overlaps y (see Breitkopf 1978:p. 231, 1.17; Simons 1987:p. 38, SCT15). We begin with parthood rather than overlap simply because parthood is the more familiar notion.

This is the entire blue region.¹²

The part of the image that stays the same



is the largest part of the image that
does not overlap the circle.

Formally, the largest part of the image that stays the same is the x such that (i) x is part of the image, (ii) x stays the same, and (iii) every part of the image that stays the same is part of x . Then given (19), we have:

- (31) The part of the image that stays the same is the x such that
- a. x is part of the image,
 - b. x does not overlap the circle, and
 - c. for all y , if y is part of the image and y does not overlap the circle, y is part of x .

Historical interlude: Leśniewski's definition of remainders.

It turns out the notion we have just defined was already formulated in the 1920s by Stanisław Leśniewski. Leśniewski defined the following notion of *relative complement* between objects.

P is the complement of the object Q with respect to R if and only if the following conditions are satisfied:

- (α) Q is a subset of the object R ;
- (β) P is the class of elements of the object R external to Q .

Leśniewski (1927–1931), translated from the Polish by Sinisi (1983:29, Definition VII).

Despite the set-theoretic terminology, in Leśniewski's system this is a purely mereological notion.¹³ Simons (1987) reconstructed Leśniewski's Mereology in

¹²As an aside, this illustrates an interesting fact about the definite article *the*. There has been a long debate about whether definite articles presuppose uniqueness. The fact that *the part of the image that stays the same* is acceptable shows that *the* does not in general presuppose uniqueness, but only given a domain of entities that do not overlap; see Casati and Varzi (1999:112), Chierchia (2010) and Kratzer (2012:168).

¹³One suspects Leśniewski deliberately adopted set-theoretic terms to show that his Mereology can do everything set theory can do, without the apparent platonism (see Simons 2020:§4.3). For some arguments that set-theoretic mereology is not enough to serve as a foundation of mathematics, see Hamkins and Kikuchi (2016).

modern terminology, offering the following definition of mereological difference.

If x and y are two individuals, then their mereological difference,

$$x - y$$

is the largest individual contained in x which has no part in common with y .

(Simons 1987:14)

That is, $x - y$ is defined by satisfying the following three properties.

- $x - y$ is part of x .
- $x - y$ does not overlap y .
- For all z , if z is part of x and z does not overlap y , z is part of $x - y$.

When the remainder $x - y$ exists, it is unique (by antisymmetry of parthood). This licences us to speak of *the* remainder $x - y$.

We can use this subtraction operation to express our findings in the circle example. When we imagine the circle a different colour, the part of the image that stays the same is the image minus the circle. Given our definition of mereological difference, we can formulate the following equation, where the grey check pattern represents the absence of colour.



Now, what does it mean for this remainder to ‘stay the same’ when we imagine the circle a different colour? Intuitively, it is for the remainder to be part of the images we imagine.



The subtraction operation also allows us to straightforwardly express our analysis of *ceteris paribus*. Given a state x , when we are asked to vary y , the *ceteris* is $x - y$, and we may define that z is *ceteris paribus* with respect to x when varying y just in case $x - y$ is part of z .

We therefore have two ways to think of what stays the same when we imagine varying a part of a moment, corresponding to two notions of *s*-variant: a plural version and a singular version. Recall the plural version that t' is an *s*-variant of t just in case every part of t that does not overlap s is part of t' (Definition (29)). We now also have a definition that operates on a single remainder directly.

$$\underbrace{x - y}_{\text{ceteris paribus}} \leq \underbrace{z}_{\text{ceteris paribus}}$$

- (32) **Singular version:** For any state s and moments t and t' ,
 t' is an s -variant of t just in case if $t - s$ exists, it is part of t' .

The proviso ‘if $t - s$ exists’ covers the case where every part of t overlaps s (in other words, t and s are inextricable) so when we remove s from t there is nothing left. For example, removing t from itself leaves nothing: $t - t$ does not exist.

This singular notion of s -variant is arguably simpler than the plural version, since to determine whether a world t' is an s -variant of t we need only check whether a single state, $t - s$, is part of t' . Under some plausible constraints on parthood, it turns out that the singular and plural definitions are equivalent. Let us now see what those constraints are.

Under what conditions do remainders exist?

Remainders can fail to exist for three reasons. Firstly, $s - t$ fails to exist when s has no part disjoint from t , as in the structures of Figure 3.5 (the two rightmost examples are from Simons 1987:27–28).

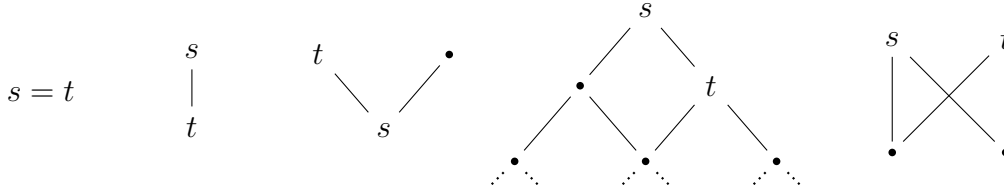
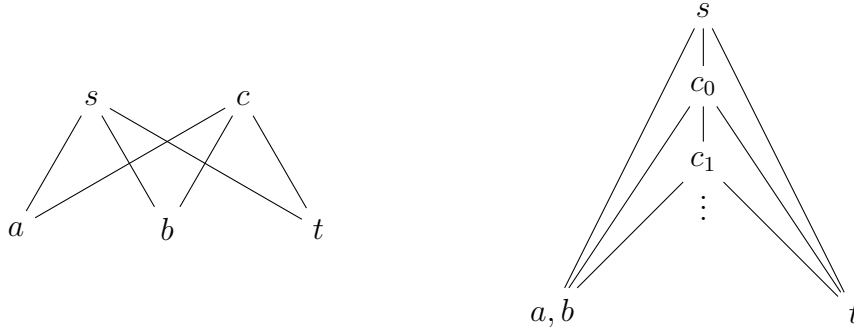


Figure 3.5: Orders where no part of s is disjoint from t .

Let us say that t is *extricable from* s just in case s has a part disjoint from t . Then in these structures, t is inextricable from s . If we try to remove t from s , intuitively there is nothing left. Our definition of remainder validates this intuition: if t is inextricable from s , $s - t$ either does not exist, or is the null state (if it exists; the null state is the state that is part of every state).

Secondly, $s - t$ can fail to exist when the order is not bounded complete.¹⁴ Bounded completeness fails when a set of states with an upper bound has no least upper bound, as in the structures of Figure 3.6 (where a , b and t are each part of c_i for every natural number i). In this structure the set $\{a, b\}$ has an upper bound but no least upper bound.

¹⁴For any partial order (S, \leq) , $s \in S$ and $T \subseteq S$, we call s is an *upper bound* of T just in case $s \leq t$ for all $t \in T$, and s a *least upper bound* of T just in case s is an upper bound of T and $s \leq b$ for every upper bound b of T . A partial order (S, \leq) is *bounded complete* just in case

Figure 3.6: $s - t$ does not exist due to failures of bounded completeness.

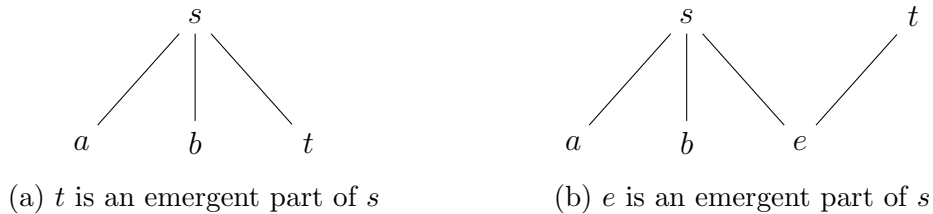
Given this structure, if we try to remove t from s , what intuitively is left over? States a and b are the states disjoint from t , so a natural answer would be that a and b are left over. Certainly, we need to remove s itself and each c -state since they contain t : keeping them around would also keep t around. So one could propose that $s - t$ is the set $\{a, b\}$; however, we are not looking for a set (an abstract object), but a state, a part of the world. The problem with these structures is that the parts of s disjoint from t do not form a state, in the sense that they do not have a fusion. (The fusion of a set of states T , denoted $\bigsqcup T$, is its least upper bound with respect to parthood.)

Thirdly and finally, $s - t$ can fail to exist due to emergent parts: states that are part of a fusion without overlapping any of the fused states.

3.3.1. DEFINITION (Emergent part). Let (S, \leq) be a partial order and $s \in S$ a state and $T \subseteq S$ a set of states. We call s an *emergent part* of $\bigsqcup T$ just in case s is part of $\bigsqcup T$ but does not overlap any $t \in T$.

We say t has an emergent part just in case some $s \in S$ is an emergent part of t , and that (S, \leq) has an emergent part just in case s is an emergent part of t for some $s, t \in S$.

Examples of emergent parts are given in Figure 3.7.

Figure 3.7: $s - t$ does not exist due to emergent parts.

every nonempty subset of S with an upper bound has a least upper bound.

Here a and b are the parts of s disjoint from t , but their fusion $a \sqcup b$ still overlaps t . Indeed, $a \sqcup b$ is s : we tried to remove t from s but are back to where we started. Even though s has parts disjoint from t , when we take their fusion, the emergent part acts as a backdoor allowing some of t to sneak back in.¹⁵

We have seen three ways the remainder $s - t$ can fail to exist: t is inextricable from s , the order is not complete, or the order has emergent parts. Are these the only ways $s - t$ can fail to exist? It turns out they are, as shown in the following facts.

3.3.2. FACT. Let (S, \leq) be a partial order and $s, t \in S$. Let $s \setminus t := \{s' \leq s : s' \text{ does not overlap } t\}$. Then $s - t$ exists if and only if $s \setminus t$ has a least upper bound with no emergent parts.

PROOF. (\Rightarrow) Suppose $s - t$ exists. Then by definition, $s - t \in s \setminus t$ and $s - t$ an upper bound of $s \setminus t$. Then $s - t \leq x$ for every upper bound x of $s \setminus t$, so $s - t$ is a least upper bound of $s \setminus t$. Hence $s - t = \bigsqcup(s \setminus t)$. To show that $\bigsqcup(s \setminus t)$ has no emergent parts, pick any $x \leq \bigsqcup(s \setminus t)$. Then $x \leq s - t \in s \setminus t$, so x overlaps an element of $s \setminus t$. Then x is not an emergent part of $\bigsqcup(s \setminus t)$.

(\Leftarrow) Suppose $\bigsqcup(s \setminus t)$ exists and has no emergent parts. We show that $s - t$ exists by showing that $\bigsqcup(s \setminus t) = s - t$; that is, $\bigsqcup(s \setminus t)$ satisfies the definition of $s - t$:

- (i) $\bigsqcup(s \setminus t)$ is part of s ,
- (ii) $\bigsqcup(s \setminus t)$ is disjoint from t ,
- (iii) every part of s disjoint from t is part of $\bigsqcup(s \setminus t)$.

(i). By construction of $s \setminus t$, $s' \leq s$ for all $s' \in s \setminus t$, i.e. s is an upper bound of $s \setminus t$. Then as $\bigsqcup(s \setminus t)$ is the least upper bound of $s \setminus t$, $\bigsqcup(s \setminus t) \leq s$. (ii) Suppose for reductio that $\bigsqcup(s \setminus t)$ overlaps t , i.e. for some u we have $u \leq \bigsqcup(s \setminus t)$ and $u \leq t$. Since $\bigsqcup(s \setminus t)$ has no emergent parts, u overlaps some $s' \in s \setminus t$, i.e. for some v , $v \leq u$ and $v \leq s'$. Then $v \leq u \leq t$, so s' overlaps t , contradicting the fact that $s' \in s \setminus t$. (iii) Pick any $s' \leq s$ disjoint from t . Then $s' \in s \setminus t$, so $s' \leq \bigsqcup(s \setminus t)$. \square

Intuitively, $s - t$ only exists when s has a part disjoint from t . What do we need to assume about parthood to guarantee this? The following result gives an answer: bounded completeness and no emergent parts.

¹⁵One motivation for assuming that the state space has no emergent parts is that, given no emergent parts, our notion of fusion (the algebraic notion, i.e. least upper bound) satisfies another popular definition of fusion; namely, that a state s is the fusion of a set of states T just in case any state overlaps s just in case it overlaps a state in T . This notion of fusion is used by Simons (1987:SD9, p. 37) and Casati and Varzi (1999:p. 46). Cotnoir (2018) calls this notion ‘Goodman fusion’ after Goodman (1951). If the state space has no emergent parts, then every fusion in our sense is a Goodman fusion. For a comparison of different notions of fusion see Hovda (2009) and Loss (2021).

3.3.3. FACT. Let (S, \leq) be a bounded complete partial order with no emergent parts. For any $s, t \in S$, $s - t$ exists if and only if s has a part disjoint from t .

PROOF. Let (S, \leq) be a bounded complete partial order with no emergent parts. (\Rightarrow) If $s - t$ exists, then by definition, $s - t$ is a part of s and disjoint from t . (\Leftarrow) If s has a part disjoint from t , then $s \setminus t$ is nonempty. Then as $s \setminus t$ is a nonempty set with an upper bound, by bounded completeness it has a least upper bound $\bigsqcup(s \setminus t)$. Then as $\bigsqcup(s \setminus t)$ has no emergent parts, by Fact 3.3.2, $s - t$ exists. \square

Intuitively we should expect the remainder $s - t$ to exist just in case s has a part disjoint from t . Fact 3.3.3 tells us that under two assumptions (completeness and no emergent parts) this expectation is guaranteed to be met.

3.3.4. FACT. Let (S, \leq) be a partial order. If (S, \leq) is bounded complete and has no emergent parts, then for any state $s \in S$ and maximal states $t, t' \in S$, the following are equivalent.

1. t' is an s -variant of t .
2. If $t - s$ exists, it is part of t' .

PROOF. $(1) \Rightarrow (2)$. Suppose (1) and that $w - s$ exists. By definition, $w - s$ is part of w and does not overlap s , so by (1), $w - s$ is part of w' . $(1) \Leftarrow (2)$. Suppose (2) and pick any part t of w that does not overlap s . Either $w - s$ exists or it does not. Suppose it does. Then as $w - s$ is the largest part of w disjoint from s , t is part of $w - s$, and as $w - s$ is part of w , by (2), $w - s$ is part of w' , so by transitivity of parthood t is part of w' . Hence (1). So suppose instead $w - s$ does not exist. Since (S, \leq) is bounded complete with no emergent parts, by Fact 3.3.3, every part of w overlaps s , so (1) is vacuously true. \square

Fact 3.3.4 tells us that assuming bounded completeness and no emergent parts, our plural and singular notions of s -variant coincide.

Now, it may seem that we have spent a great deal of time stating the obvious. When we are asked to imagine the circle with a different colour, it is obvious that the blue part stays the same. But our discussion has revealed a general principle of hypothetical reasoning: a state stays the same just in case it does not overlap a part we wish to vary. The next section demonstrates the generality of this proposal.

3.4 The varieties of parthood

We have seen that we can use overlap to account for what stays the same when imagine a state changed. Overlap itself can be defined in terms of parthood: two things overlap just in case something is part of both of them. Now, our proposal in (30) does not tell us how the notion of parthood is to be understood. So far we

have only considered one kind of parthood: parthood in space. This is perhaps the paradigmatic case, but it is not the only way to understand the notion.

We speak, for example, of the parts of an organisation. University departments are part of faculties, which are part of universities. When we imagine changing the state of a faculty (say, by imagining hiring new people at the faculty) we do not keep the state of the whole university fixed. This parallels how when we imagine changing a part of an image we do not keep the whole image fixed. And if two faculties overlap – say, the logic department is part of the faculty of humanities and the faculty of science, as it is in Amsterdam – then imagining a change to one of the faculties can result in imagining a change to the other; for if one imagines changing the state of the humanities faculty by changing the logic department, that will change the state of the science faculty too. This parallels how when change a part of an image, we do not keep fixed any part of the image that overlaps the part we wish to change.

Now, while each faculty may have a physical location, the faculty is not identical to its location. A faculty is part of the university, but this is not a spatial notion of parthood. It is parthood not in physical space, but in conceptual space. In section 3.3 we introduced our proposal with a spacial notion of parthood in mind, but now we can appreciate how it gives the right results when applied to parthood broadly construed. For example, applied to the sense of parthood in which departments are part of universities, our proposal in (30) still gives desirable results: when we imagine changing the state of a university department s , the state of a university department t stays the same just in case t does not overlap s in conceptual space.

We also speak of an object having its properties as parts. The fact that Socrates is a philosopher is part of who he is. Here are some naturally occurring examples of parthood used to describe one's properties.

- (33) a. Alison was a beautiful and caring person who I am incredibly proud of, and that is the part of her I want everyone to remember.
 b. The part of her we can't see in airbrushed photos is her challenging personal history.¹⁶

Or consider the following object, which is made of gold.

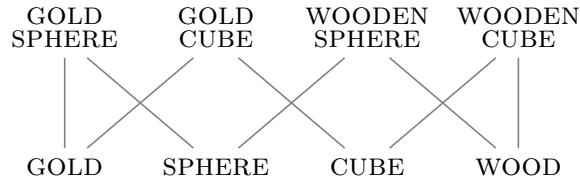


Now imagine it made of wood. What comes to mind? As a matter of fact, we tend to imagine it still a sphere. When changing the material, we keep the shape fixed. We can represent this fact on the present proposal if the properties of an object are part of it. For example, the goldness of a gold sphere is part of it, and

¹⁶Sources: (33a) *Hull Daily Mail*, 23 November 2021; (33b) *Unstoppableteen.com*.

an object's material does not overlap its shape in conceptual space. This is not parthood in space, since an object's shape and material, if they are located in space at all, presumably have the same spatial location.¹⁷

Here is a toy state space representing material and shape, featuring only two materials (gold and wood) and two shapes (sphere and cube) together with their possible combinations.



Given this state space, the properties of the object being gold and being spherical do not overlap. So when we vary SPHERE, according to our proposal, GOLD stays the same, as desired.

Note that the states in the figure above do not represent universals, such as the property (or Platonic form) of goldness itself, but instead the goldness of a particular object – what in the metaphysics literature might be called a thick particular (Armstrong 1989), state of affairs (Armstrong 1997) or trope (Williams 1953, Campbell 1981, Moltmann 2007). Two objects can share the same property while we imagine changing one but not the other. For example, given two gold spheres, *a* and *b*, imagine *a* is made of wood. Intuitively *b* is still made of gold.

If we want to start with a state space where each state represents a universal property, a simple way to represent the properties of particular objects would be to attach names to the states (the name represents what in the metaphysics literature is called a *thin particular*). This allows us to distinguish, say, the goldness of object *a* from the goldness of object *b*, as in the state space of Figure 3.8 (for simplicity, we omit further combinations of states, such as the state of *a* being gold and *b* a sphere).

Since GOLD_a and GOLD_b do not overlap, our proposal predicts that when varying one, we keep the other fixed, which is intuitively the right result.

It is an interesting question how we construct a state space from experience; for example, how we know that a gold sphere's properties of being gold and being a sphere do not overlap. This is a deep question that we will not try to answer here. It is presumably given by the structure of the world and/or our cognition. The state space is a primitive of the theory. The present paper is not intended to

¹⁷There is a longstanding idea in metaphysics that the properties of an object are in some sense part of it. This view often goes by the name *constituent ontology*. For an overview see Loux (2005, 2012). For discussion see van Inwagen (2011), Forrest (2013), Olson (2017), and Yang (2018). Armstrong (1986, 1988) argues that exemplification is a non-mereological mode of composition. In the text I use 'parthood' to cover both ordinary cases of parthood (the sense in which, for example, one's hands are part of one's body) and exemplification (the sense in which the redness of a rose is part of it).

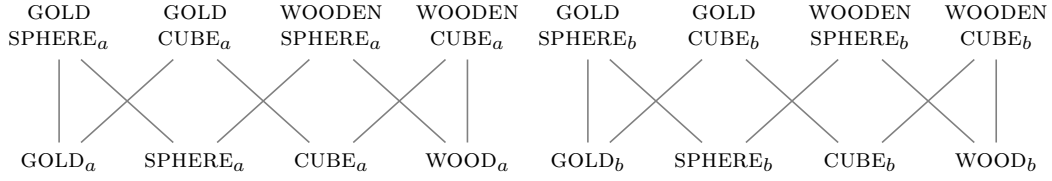


Figure 3.8

propose a particular mereological structure of (our cognition of) the world, but to propose the existence of a correlation between mereological structure and how we construct hypothetical scenarios. In this way, the present paper contributes to – without, of course, answering – the age-old question: What structure must our experience have for us to be able to reason in the way that we do?

3.4.1 Imagination in an atomless world

It may be tempting to think that to represent how we imagine changing a part of the world, we must represent the world at a certain level of granularity, as a combination of parts that are taken to be atomic.¹⁸ This, however, is not required on the present proposal.

Consider a glass of water. Now imagine the glass filled with a different material. Intuitively, we do not keep fixed the matter that makes up the water: the molecules, the atoms, the quarks and so on. These are liable to change. Now, it could be that the world is non-atomic, in the sense that every part of the world has proper parts. While contemporary physics suggests that there is a smallest unit of length – the Planck length, see Mead (1964) – there is no reason why our logical models should decide this issue, forcing atomic representations upon us. We can reason hypothetically without assuming atomism. Suppose, for the sake of argument, that there are no atoms. When we imagine a glass of water filled with a different material, intuitively we allow all parts of the water to vary. Even though the water is assumed to have infinitely many parts, that is no obstacle to imagining that whatever is part of the water is allowed to vary.

This is predicted on the present approach, since whatever is part of the water overlaps the water, and is therefore allowed to vary when we imagine the glass not containing water. We might, for example assume the mereological structure in Figure 3.9 (which is of course greatly simplified; for example, it does not show the interactions between the parts of the molecule, such as the bonds).

¹⁸Pollock (1976), for example, uses the notion of ‘simple propositions’ in his semantics of counterfactuals. See Kratzer (2012:74) for some problems with Pollock’s account. We discuss Kratzer’s approach in sections 4.3.1 and 4.4.1.

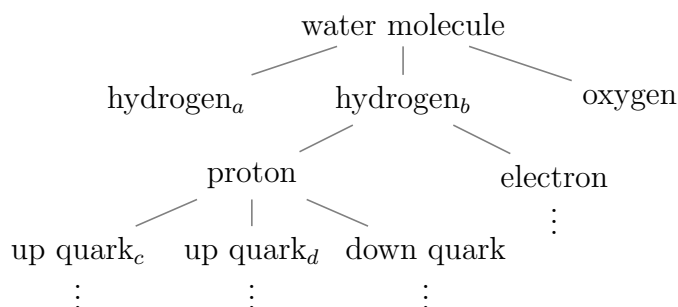


Figure 3.9

3.4.2 Dependence between properties

The present approach can also handle dependence between properties. To illustrate, consider a ring made entirely of pure gold. Now imagine the ring a different colour. Is the ring still made of entirely of pure gold? Intuitively not. The material must change too. We do not imagine the ring being, say, both blue and made entirely of gold, since such a material does not exist. As every florist and painter knows, colour and material are dependent: roses are red, violets are blue, as is lapis lazuli, Egyptian mummies are brown, gold is yellow (while we do have names for white gold, rose gold, and even blue gold, these are not pure gold but alloys of gold and another material). Other materials, like plastic and paper, come in all colours. When we imagine a blue piece of paper with a different colour, intuitively we still imagine a piece of paper.¹⁹

We can capture these facts on the present proposal. Let us assume for simplicity that gold is yellow (i.e. we ignore gold's shininess). We can represent the parthood relation between some materials and colours as follows.

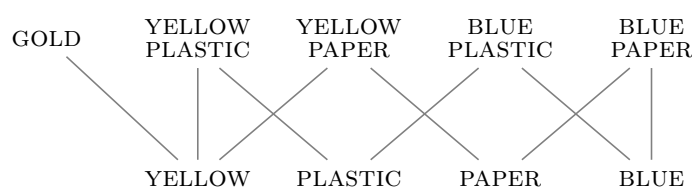


Figure 3.10

According to this state space, part of what it is for something to be gold is for it to be yellow – as Kant put it, “my concept of gold ... includes that this body is yellow” (*Prolegomena* 4:267) – just as part of what it is for something to be blue paper is for it to be blue. On the present proposal, since YELLOW is part

¹⁹It was arguably dependencies such as these that led Wittgenstein to abandon the logical atomism of his *Tractatus* (see Pears 1981, Jacquette 1990). Another example of a dependence between states, from Wittgenstein (1977), is the impossibility of transparent white.

of GOLD, YELLOW overlaps GOLD, so when we imagine a change to the colour of a gold ring we do not fix the fact that it is made of gold. In contrast, when we imagine, say, a blue piece of paper being a different colour, since PAPER does not overlap BLUE, we fix the fact that it is made of paper. These are intuitively the right results.

Thus state spaces can represent the fact properties are in some cases independent and in other cases dependent. For example, colour and material are independent when the material is plastic but dependent when the material is gold.

Now, why is it important to be able to represent dependence between properties? There are some cases illustrating that when we construct hypothetical scenarios, we can take dependence between properties into account. Here is one such case. Imagine a machine – in a recycling plant, say – that sorts objects using two sensors. It has a material detector that determines an object’s material and a camera that determines its colour. (In case it matters, let us suppose that the sensors perform their tests at the exact same time.) The two sensors are programmed independently: one selects what materials for the material detector to accept and separately, what colours for the camera to accept. The two sensors each send their verdict of **accept** or **reject** to a central computer, which controls whether to place the object in the accept bin or the reject bin. If both sensors return **accept**, the object is placed in the accept bin, otherwise it is placed in the reject bin.

Suppose we set the material checker to accept plastic and the camera to accept blue objects. We give the sorter Object A: a yellow piece of plastic. It is rejected (see Figure 3.11).

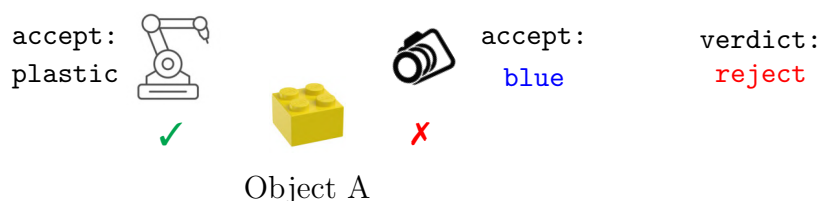


Figure 3.11

Asked about what happened, we could reply:

(34) If object A had been blue, it would have been accepted.

(34) is a perfectly natural thing to say in this scenario. This suggests that we can vary object A’s colour while fixing its material. When we imagine a yellow piece of plastic blue, we imagine a blue piece of plastic.

Now suppose we reset the material checker to accept pure gold. We leave the camera as is, set to accept blue objects. We give the sorter object B: a pure gold

ring. It is also rejected (see Figure 3.12).

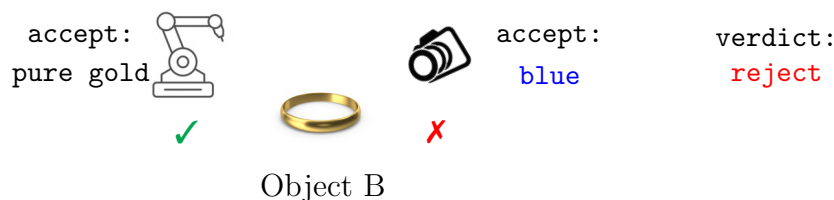


Figure 3.12

Consider (35) in this context.

(35) If object B had been blue, it would have been accepted.

(35) seems unacceptable. There is a clear contrast in acceptability between (34) and (35). Intuitively, if object B had been blue, it still would have been rejected since in that case it would have failed the material test. Indeed, on reflection, with the material checker set to gold and the camera set to blue, any object whatsoever would be rejected.

Of course, we could expand our imaginative horizons to consider blue gold. If one saw a newspaper article titled *Scientists create blue form of pure gold!* one might be inclined to believe it. In our actual world, however, we know blue gold is impossible (indeed, this is due to relativity; see Pyykkö and Desclaux 1979). Under the currently true assumption that blue pure gold does not exist, it hard to see how (35) could hope to be true.

One may respond that such a sorting machine is implausible. One may prefer a more sophisticated machine, featuring automatic dependence recognition, whereby, for example, if one sets to material detector to accept gold, the camera is forced to accept yellow, and if one sets camera to accept blue, the material detector cannot accept gold. However, it is perfectly possible to build a ‘budget’ sorting machine of the kind above, without automatic dependence recognition, where the choice of which material and colour to accept are set independently.

These examples show that to capture hypothetical reasoning, we need a framework that can represent properties (such as colour and material) as dependent in some cases and independent in others. It is hard to see how we could do this if we represented properties as random variables, where the set of possible scenarios corresponds to the set of assignments of values to the variables. This conception is ubiquitous in formal modelling, such as probability theory, Bayesian networks and structural causal models, the atomic sentences of logic, Euclidean space, the phase spaces of dynamical systems theory, and so on.

In light of this tradition, it is tempting to think that when we reason hypothetically, we also represent the scenario in question using random variables. Let us see how one framework that uses random variables handles dependence

between properties: the causal graphs of Spirtes, Glymour, and Scheines (1993) and Pearl (2000). In that framework, the nodes of the graph represent variables, and directed edges represent dependencies between variables. Changing the value of one variable (called *intervening* on it) keeps fixed the values of all of its non-descendants. That is, when varying variable X , the value of variable Y is held fixed if there is no directed path from X to Y in the causal graph. Figure 3.13 depicts a causal graph of the scenario containing object B.

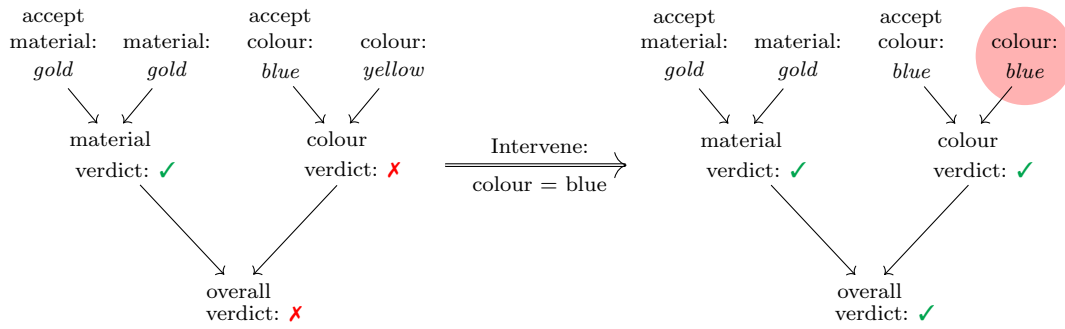


Figure 3.13

Pearl (2000: chapter 7) presents a semantics of counterfactuals where $A > C$ is true in a structural causal model M and context u just in case C is true in the model that results from intervening on M and u to make A true. When we intervene to set the colour variable of object B to blue, we find that the object is accepted. Given this model, then, Pearl's semantics incorrectly predicts (35) to be true.

This model is, to my mind, the most natural structural causal model of the scenario above. Stepping back, one may wonder whether *any* structural causal model can the dependence between properties that we have discussed. We will not conclusively answer this question, but there is reason to think prospects are bleak. The fact that we judge (34) true shows that when we imagine a piece of yellow plastic being blue, we intuitively fix its material. We can capture this by representing colour and material as separate variables, with no directed path from colour to material. But the fact that we do not judge (35) to be true shows that when we imagine a piece of gold being a different colour, intuitively we do not fix its material. To represent this using causal graphs now we need a directed path from colour to material. Given this conflict, it is unclear how to proceed using structural causal models.

Let us consider two responses to this problem on behalf of structural causal models. The first is to add a restriction to possible variable assignments, or alternatively, to allowed interventions.²⁰ The question for this proposal is what

²⁰For example, Beckers and Halpern (2019: Definition 2.2) distinguish between allowed and forbidden interventions, though not to account for dependence between properties.

to do when an intervention takes us to an impossible assignment, such as one with blue gold. Intuitively, imagining a gold object being blue does not result in a breakdown of imagination: we simply imagine the material changing too, while fixing other facts (such as the object's shape). Ruling out certain assignments or interventions does not reflect the systematic nature of the imagination – where some facts stay the same and others vary – in cases where the properties in question are dependent, such as colour and material.

A second response on behalf of structural causal models is to represent colour and material as a single variable. For this proposal to work, when one wishes to intervene to change an object's colour, as in (34) and (35), one needs to add a rule stating how interventions on colour are translated into interventions on the single colour/material variable. To capture the acceptability of (34) and the unacceptability of (35), one needs the rule to fix material when interpreting (34) but not (35). Structural causal models do not specify how to determine this rule.

No such difficulties arise on the present proposal. We simply use a state space where YELLOW overlaps GOLD but does not overlap PLASTIC, such as the state space in Figure 3.10. Given our proposal that we fix those states not overlapping the state we wish to vary, this is enough to predict that when we vary YELLOW for a plastic object, we fix the fact that it is plastic, but when we vary YELLOW for a gold object we do not fix the fact that it is gold.

3.4.3 Imagination without variables

In this section we show that partial orders are strictly more general than random variables, in the sense we can represent every set of random variables as a state space, but not vice versa.

From variables to state spaces. Translating from variables to partial orders is straightforward. We take each state to be an assignment of values to variables.

3.4.1. DEFINITION (Variable space). Let V be a nonempty set, where each $X \in V$ is associated with a range $R(X)$ taken from a set *Values*. We say an *assignment of values to some variables in V* is a function $s : U \rightarrow \text{Values}$ such that U is a nonempty subset of V and $s(X) \in R(X)$ for all $X \in U$.

We say (S, \leq) is *generated by V* just in case

$$\begin{aligned} S &= \text{the set of assignments of values to some variables in } V \\ \leq &= \subseteq \end{aligned}$$

We call a partial order a *variable space* iff it is generated by a set of variables.

Figure 3.14 gives an example of a set of variables and the state space it generates. In the Figure we let $x\bar{y}$, for example, stand for the function with domain $\{X, Y\}$ mapping X to x and Y to \bar{y} .

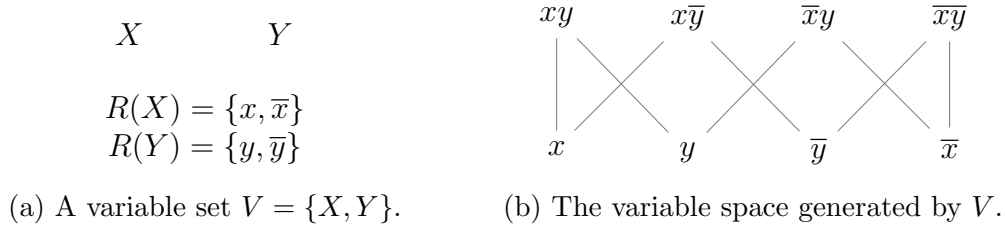


Figure 3.14

Variable spaces are especially well-behaved, enjoying all of the desirable properties of parthood we consider in this essay.

3.4.2. FACT. Every variable space is a partial order, a world-space, bounded complete, and has no emergent parts.

PROOF. Let (S, \leq) be any variable space generated by variables V . \leq is a partial order since inclusion is a partial order. World-space: The set of maximal elements of (S, \leq) is the set of total assignments, $\{s : V \rightarrow \text{Values} : s(X) \in R(X) \text{ for all } X \in V\}$. Clearly, every state is a subset of a maxima element. Bounded completeness is immediate. No emergent parts: pick any state s and set of states T whose fusion exists. Suppose $s \leq \sqcup T$. Then $s \subseteq \sqcup T$. Then clearly, for some variable $X \in V$ and value x of X , $\{(X, x)\}$ is a subset of s and t , for some $t \in T$. So s overlaps a state in T . \square

This shows that the variable sets of Spirtes, Glymour, and Scheines (1993) and Pearl (2000) are a special case of the state spaces we consider here.

The translation from variable sets to state spaces gives us another perspective on what goes wrong when we try to represent dependence between properties using random variables. Suppose we have a colour variable whose possible values are yellow and blue, and a material variable whose possible values are gold, plastic and paper. These variables generate the state space in Figure 3.15.

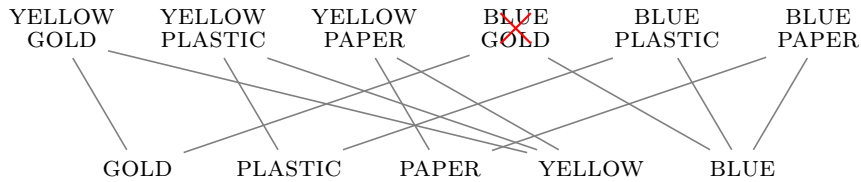


Figure 3.15

If we keep the blue gold state, we predict that when we imagine a pure gold object blue, we imagine it made of blue pure gold. This does not capture the fact that when imagine the object blue, we allow its material to vary. If we remove the blue gold state, we face a different problem: the yellow and gold states do

not overlap. Given a gold ring, the current state is YELLOW GOLD. In this state space, when we remove its yellowness, given that gold and yellow do not overlap, its goldness remains. But every object containing the GOLD state is yellow: we want to vary the object's colour, but its goldness holds us back. Thus according to this state space we cannot imagine a gold object being blue. However, as we saw above, we can indeed interpret a sentence such as (35). When we do so, we keep some properties of the object (such as its shape) but do not keep its material – something random variables cannot represent.

State spaces without a variable structure. Let us define what it means for a state space to have a variable structure, in the sense that it can be represented as the possible assignments of values to variables. We will take an abstract approach, allowing a state space to have a variable structure even if the variables only represent the state space at a certain level of abstraction. For it may be that we have a complex state space but only want to describe certain aspects of it using variables, ignoring others.

We assume that state spaces are moment spaces, in the sense that every state is part of a maximal state with respect to parthood (called a *moment*). We define a *variable* to be a set of states X where every moment contains exactly one state in X . Then given a variable X , we can think of an element x of X as a possible value of X , and can define the sentence $X = x$ to be true at a moment m just in case x is part of m . The requirement that every moment contain a unique state in X ensures that every moment assigns a unique value to each variable.

To illustrate, consider a state space representing some colours and shapes, depicted in Figure 3.16. According to our definition of variable, this state space has three variables: the colours, the shapes and the moments.

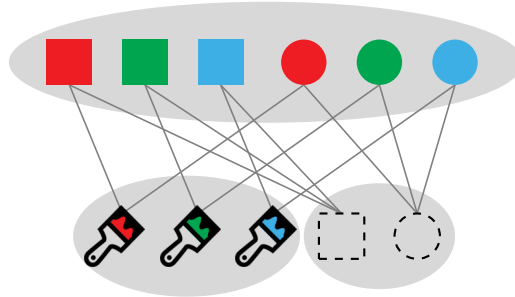


Figure 3.16

Given a set of variables V , an *assignment of values to V* is a function assigning to each variable X in V an element of X .

Intuitively, a set of variables represents a set of scenarios just in case it satisfies the following two conditions.

Orthogonality. Every variable assignment is consistent. For every assignment

v of values to V , there is a moment m such that every state in v is part of m .

Expressiveness. Every moment can be expressed as a variable assignment. For every moment m , there is an assignment v of values to V such that every state in v is part of m .

The state space above has two variable sets satisfying orthogonality and expressiveness: $\{\text{COLOURS}, \text{SHAPES}\}$ and $\{\text{MOMENTS}\}$. Let us call a variable set *trivial* just in case it contains a single element. Note that for any moment space (that is, partial order where every state is part of a maximal state) the trivial variable set containing only its maximal elements satisfies orthogonality and expressiveness. So let us say that a state space has a *variable structure* just in case it has non-trivial variable set satisfying orthogonality and expressiveness.

The question naturally arises whether, to reason hypothetically, one must have in mind a state space with a variable structure. The answer is ‘no’. Figure 3.10 gave a state space representing the dependence between colour and material. One can prove that this state space does not have a variable structure.

3.4.3. FACT. The state space in Figure 3.10 does not have a variable structure.

PROOF. Suppose for reductio that the state space in Figure 3.10 has a nontrivial variable set V . Since GOLD is a world, by expressiveness, there is an assignment v of values to V such that every state in v is part of GOLD. Either $\text{GOLD} \in v$ or not.

Suppose $\text{GOLD} \in v$. Then $\text{GOLD} \in X$ for some variable $X \in V$. Since V is non-trivial, $Y \in V$ for some variable $Y \neq X$. By orthogonality, GOLD is consistent with every value of Y (that is, for every $y \in Y$ there is a world containing both GOLD and y). Now, GOLD is only consistent with itself and YELLOW, so $Y \subseteq \{\text{GOLD}, \text{YELLOW}\}$. Since Y is a variable, every world contains exactly one state in Y , contradicting the fact that, for example, BLUE PAPER contains neither GOLD nor YELLOW.

Suppose $\text{GOLD} \notin v$. By non-triviality, there are distinct variables $X, Y \in V$. So there are states $x \in X$ and $y \in Y$ such that $x, y \in v$ and $x, y \leq \text{GOLD}$. The only possible choices for x and y are GOLD and YELLOW, i.e. $x, y \in \{\text{GOLD}, \text{YELLOW}\}$. Now since $X \neq Y$, by Fact 3.4.4, X and Y are disjoint, so $x \neq y$. Then $x = \text{GOLD}$ or $y = \text{GOLD}$. Then as $x, y \in v$, $\text{GOLD} \in v$, contradicting $\text{GOLD} \notin v$. \square

3.4.4. FACT. For any variable set V satisfying orthogonality, all distinct variables X and Y in V are disjoint.

PROOF. Let X and Y be distinct variables. Then $x \in X \setminus Y$ for some state x , or $y \in Y \setminus X$ for some state y . Suppose w.l.o.g. that $x \in X \setminus Y$. Now suppose for reductio that X and Y are not disjoint, i.e. $z \in X \cap Y$ for some state z . By orthogonality, there is a world w containing x and z , but then w contains two

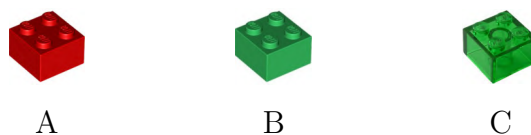
values of X , contradicting the fact that X is a variable. \square

I conclude that we do not need random variables to model hypothetical reasoning.

3.5 Case study: state spaces of colour

The present way of representing information via parthood (rather than random variables) may be unfamiliar to some who are engaged in formal modelling. Let us therefore consider one further case study to illustrate the expressive power of state spaces. In this section we propose a state space of colour. The relationship between colour concepts has captivated philosophers across centuries, such as Newton, Goethe and Wittgenstein. For example, Wittgenstein (1977:5, 43) wondered whether there is “a logic of colour concepts”. In this section we will see that, indeed, state spaces allow us to model hypothetical reasoning about colour.²¹

One example of dependence between properties concerns colour and opacity. To illustrate, consider the bricks below.²²



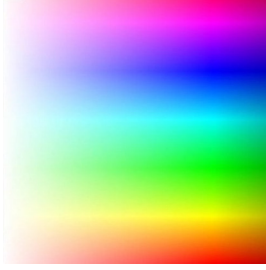
Imagine brick A green. What do you imagine? We tend to imagine brick B rather than brick C. We keep the opacity fixed while varying the colour. Now imagine brick C opaque. What do you imagine? We tend to imagine brick B rather than brick A. We keep the colour fixed while varying the opacity.

What state space would one need to derive these facts on the present approach? Let us restrict attention to the two properties of interest: colour and opacity. A first, naïve thought is our state space represents each combination of colour and opacity by an atomic state, with no parthood relations between them. Of course, such a state space does not have enough structure. The states does not distinguish colour and opacity, whereas our intuitive conceptual space does. We recognise, for example, that there is something in common between bricks A and B below (their opacity) that is not shared by brick C, and there is something in common between bricks B and C (their colour) that is not shared by brick A.

²¹The conceptual spaces of Gärdenfors (2000, 2014) are another way to represent colour properties. Gärdenfors does not provide a way to model hypothetical reasoning (e.g. imagining changing the properties of an object) which is our main concern here, so we will not consider conceptual spaces here. It is likely, however, that conceptual spaces determine an overlap relation between states, which is all we need to apply the present approach to varying a state.

²²For colourblind readers, and those reading this printed in black and white, brick A is red while bricks B and C are green; bricks A and B are opaque while brick C is semi-transparent.

Recall the rule we proposed above: when varying state x , keep state y fixed just in case x and y do not overlap. For this rule to give the results above, we need the colour and opacity of an object to be disjoint parts of it. This is not true in a simple state space where each combination of colour and opacity is represented by an atomic state. In this state space, for any state s whatsoever, every state is an s -variant of every state. So if we input this state space into our proposal in (30), we would predict that when we imagine brick C opaque, bricks A and B are both among the things we imagine.



This suggests that our intuitive conceptual space represents the colour and opacity of an object along different dimensions. One way to represent this is in the image on the left, where the x axis corresponds to opacity and the y axis to colour (for simplicity we ignore the other aspects of colour: saturation and value). Photo editing software often uses colour models like this. By representing opacity and colour along different dimensions, this colour space allows one to represent changing one while fixing the other.

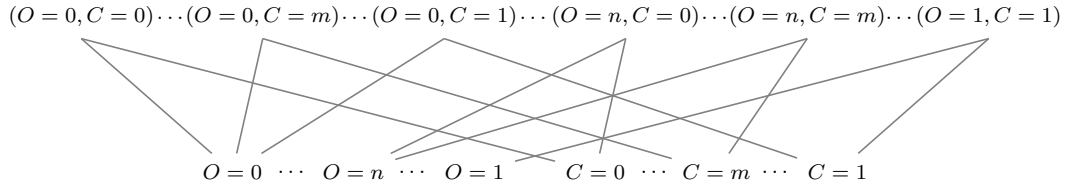
Changing opacity while fixing colour corresponds to a horizontal move in colour space. Changing colour while fixing opacity corresponds to a vertical move in colour space.

State spaces (states ordered by parthood) can also represent opacity and colour along different dimensions. Let us introduce a state for each opacity value and each colour value, together with a state for each combination of opacity and colour. One state s is part of another state t just in case every value assigned by s is also assigned by t . Formally, let O and C be variables representing opacity and colour, respectively, taking values from 0 to 1 (this range is, of course, arbitrary).

$$\text{opacities} = \{O = n : 0 \leq n \leq 1\} \quad \text{colours} = \{C = n : 0 \leq n \leq 1\}$$

Our states and parthood relation are then given as follows.

$$\begin{aligned} \text{states} &= \text{opacities} \cup \text{colours} \cup (\text{opacities} \times \text{colours}) \\ s \text{ is part of } t &\text{ iff } s \text{ is a subsequence of } t \end{aligned}$$



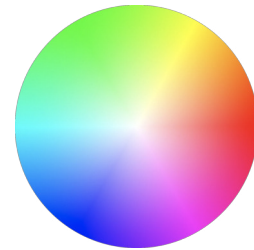
There is a clear correspondence between this state space and the colour square above. The atomic states correspond to horizontal and vertical lines in colour space, the composite states correspond to points.

When we apply our proposal in (30) to this state space, we find that when we vary colour we fix opacity, and vice versa, when we vary opacity we fix colour. This is because the opacity and colour of an object do not overlap. In general, then, given a state space that represents colour and opacity of an object as disjoint parts of it, our proposal gives the intuitively correct results.

There is, however, something strange about this state space. Consider the states with zero opacity and a colour value: the points lying on the leftmost edge of the colour square. What it would mean, say, for something to be fully transparent and blue? Or fully transparent and red? What colour is a completely clear window? This sounds like a trick question. Intuitively a perfectly transparent object does not have a colour. Nothing can be both have a colour and be fully transparent at the same time.

The colour square and state space above represent opacity and colour as completely independent, in the sense that every combination of opacity and colour is possible – even the combinations of zero opacity with any colour. In contrast, according to our intuitive concept of opacity and colour, these properties are in part dependent: if something is fully transparent, it does not have a colour. Our concept of colour simply does not apply to fully transparent things. This is a partial dependence since it is confined to the zero opacity values. Every combination of nonzero opacity and colour is still intuitively possible.

Our colour square and state space above contain ‘phantom states’, such as fully transparent green and fully transparent blue: states that intuitively do not exist. One way to fix this state space is to collapse all points with zero opacity and a colour to a single point, as in the image on the right. Mathematically, we can model this colour space as a vector space with magnitude representing opacity and direction colour. The fact that fully transparent objects do not have a colour corresponds to the fact that the zero vector does not have a direction.

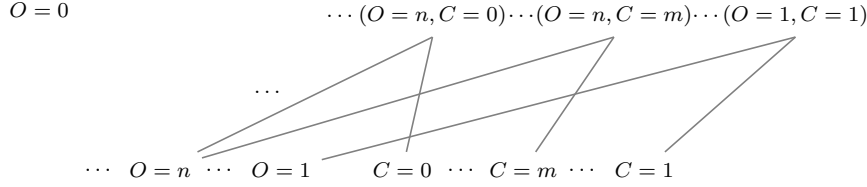


This discussion of zero opacity reveals something quite remarkable about conceptual space in general, what we may call *variable dimensionality*. In formal modelling it is common to represent a scenario as a point in some multidimensional space – an approach that assumes fixed dimensionality. But our conceptual space of opacity and colour does not have a fixed number of dimensions. It is in part one-dimensional and in part two-dimensional. When an object has some opacity the colour dimension is active, so to speak, but when it is fully transparent the colour dimension falls away.

State spaces handle variable dimensionality with ease. Given the state space above, to capture that fact that fully transparent objects do not have a colour we, quite naturally, remove all combinations of zero opacity and colour, keeping

parthood between the remaining states the same.

$$states' = states \setminus \{(O = 0, C = m) : 0 \leq m \leq 1\}$$



We end our discussion of colour with an analogy from geometry. The move from random variables to state spaces is analogous to the move from Euclidean to non-Euclidean geometry. Euclid's parallel postulate, reformulated by Playfair, says that for every line and point not on that line, there is exactly one line parallel to the first. In our colour models, points correspond to maximal states and lines to sets of points differing in one dimension; for example, $(O = .8, C = .3)$ is a point and $\{(O = n, C = .3) : 0 \leq n \leq 1\}$ a line. As usual, two lines are parallel just in case they do not share a point.

The first colour model we considered, with a state for each combination of colour and opacity, is Euclidean, satisfying the parallel postulate (see Figure 3.17).²³ However, the parallel postulate fails in our second colour model, where we removed the colour–opacity pairs with zero opacity. For any line of nonzero opacity, $l = \{(O = n, C = m) : 0 \leq m \leq 1\}$ the zero opacity point $O = 0$ is not on l , but there is no line parallel to l that contains the point.

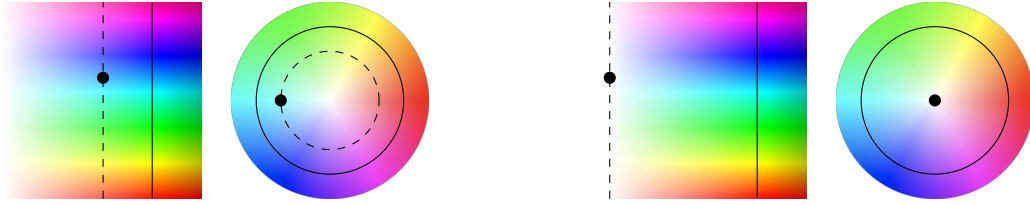


Figure 3.17: In the square, for every line l and point not on l , there is a line on the point parallel to l . In the circle, this fails at the zero opacity point.

Just as non-Euclidean geometry allows for dependence between dimensions (such as spacetime curvature in relativity), state spaces allow for dependence between properties. As we have seen, we need to represent dependence between

²³To show this, pick any line $l = \{(a, y) : y \in Y\}$ and point (b, c) not on l , i.e. $a \neq b$. Then l is parallel to $l' = \{(b, y) : y \in Y\}$, since $a \neq b$. And l' is the only line containing the point parallel to l since any line l'' is of the form $\{(x, d) : x \in X\}$ or $\{(e, y) : y \in Y\}$. If the former then l and l'' intersect at (a, d) so are not parallel; if the latter and l'' contains the point (b, c) , then $b = e$, so $l' = l''$.

properties to correctly model hypothetical reasoning. And just as non-Euclidean geometry opened up new horizons in mathematics and physics, one hopes the ability to represent dependence between properties will open up new horizons in our understanding of hypothetical reasoning.

3.6 Hypothetical scenarios raised by a sentence

We require a scheme that describes the structure of senses [meanings] in the decompositional way that chemical diagrams describe the molecular structure of compounds.







— Jerrold J. Katz, ‘Common sense in semantics’ (1982)

We have proposed a rule to capture what stays the same and what we allow to vary when we imagine changing a part of the world. We began with a simple case, changing a part of an image, and saw that what we imagine is captured by the rule that when we imagine changing a region in space, we fix every region that does not overlap the region we are changing. We then saw how this proposal makes correct predictions not only when varying a region in space, but in any case where overlap relations can be determined, such as varying a part of an organisation or the properties of an object. Now, this by itself is not enough to capture what hypothetical alternatives we consider when we interpret a causal claim or a conditional such as (1).

- (1) a. The light is off because switch A is down.
- b. If switch A were up, the light would be on.

In this case we are not given a state to vary but a sentence, which we use as a launchpad to construct hypothetical scenarios. Let us therefore turn to how sentences, rather than states, raise hypothetical scenarios.

We interpret a conditional antecedent or a *because* clause, intuitively, we identify a part of the world that needs to change, and imagine changing that. For example, given the set up of the switches in Figure 3.1 and the antecedent *if switch A were up*, we imagine changing the position of switch A. We can represent the possible positions of the switches as a state space, given in Figure 3.18, where the lines represent parthood.

Above we proposed that when we imagine changing a part of the world, a part of the world stays the same just in case it does not overlap the part that we imagine changing. In the state space in Figure 3.18,  does not overlap , so our proposal predicts that when we imagine changing one, the other stays the same. Put in terms of remainders from section 3.3.3, we have   −  = .

This by itself does not tell us what we imagine changing when we interpret conditionals or *because* claims such as those in (1). To extend this proposal to capture what scenarios we imagine when we interpret a conditional antecedent

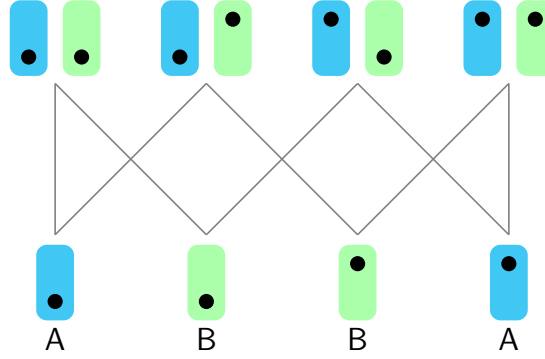


Figure 3.18

or *because* clause, we need to relate states and sentences, whereby s is related to sentence A just in case s is among the states we imagine changing when we are given a conditional antecedent or *because* clause A .

Let us first do this in a schematic way, without committing to a particular proposal on how to relate states to sentences. To do this, let us follow an idea by Ciardelli, Zhang, and Champollion (2018) in distinguishing between a sentence's 'foreground' and 'background'. We define the foreground of a sentence A to be the states we allow to vary when we interpret a conditional antecedent or *because*-clause A . Intuitively this is the set of states that A is in some sense 'about'.²⁴ Then the background of A as the set of states disjoint from every state in A 's foreground. We can then define that a moment t' is an A -variant of a moment t just in case every part of t in the background of A is part of t' .

3.6.1. DEFINITION (A -variant). Let *about* be a relation between sentences and states. For any moments t, t' and sentence A , let us call t' an A -variant of t just in case every part of t that is disjoint from every state A is about is part of t' :

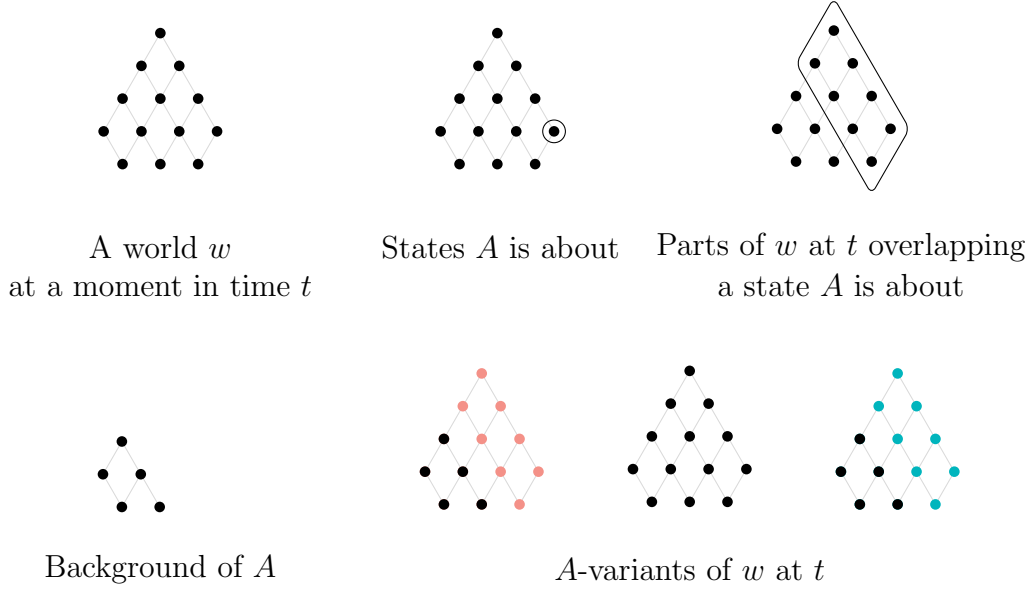
$$t' \text{ is an } A\text{-variant of } t \quad \text{iff} \quad \forall s \leq t ((\forall u \text{ } A \text{ is about } u \text{ and } s \not\leq u) \Rightarrow s \leq t').$$

This definition is illustrated in Figure 3.19.

3.6.1 A single-state definition of sentence variants

In section 3.3.3 we gave a single-state definition of state variants, whereby t' is an s -variant of t just in case if the remainder $t - s$ exists, it is part of t' . We can apply the same treatment to the definition of sentence variants.

²⁴Note that the foreground function only depends on the sentence in question – it does not depend on the world of evaluation. This is a deliberate choice. If the foreground function also depended on the world of evaluation, we would in principle allow the a state to be part of two worlds w and w' , and in the foreground of A at w but not at w' . Since additional expressive power appears unnecessary, so at present we do not allow it.

Figure 3.19: Steps to construct the A -variants of a moment.

For any sentence A and moment t , define the A -parts of t to be the parts of t that are part of a state A is about. Let us also take the A -part of t to be the fusion (i.e. least upper bound) of its A -parts, and let us abbreviate $t - A\text{-part}(t)$ as $t - A$. We may call $t - A$ the background of A at t .

$$\begin{aligned}
 A\text{-parts}(t) &= \{s \leq t : \exists u (A \text{ is about } u \text{ and } s \leq u)\} \\
 A\text{-part}(t) &= \bigsqcup A\text{-parts}(t) \\
 t - A &= t - A\text{-part}(t)
 \end{aligned}$$

3.6.2. FACT. Assume bounded completeness and no emergent parts. Then for any moments t, t' and sentence A , the following are equivalent.

1. t' is an A -variant of t .
2. If $t - A$ exists, $t - A$ is part of t' .

PROOF. Recall from section 3.3.3 that $t - A$ is the largest part of t that does not overlap $A\text{-part}(t)$.

(1) \Rightarrow (2). Let t' be an A -variant of t and suppose $t - A$ exists. Note that $t - A$ does not overlap a state A is about. For suppose for reductio that it did. Then $s \leq t - A\text{-part}(t)$ and $s \leq u$ for some state u that A is about. Then $s \leq t - A \leq t$, so by definition of $A\text{-part}(t)$, $s \leq A\text{-part}(t)$. But then $t - A$ overlaps $A\text{-part}(t)$, contradicting the definition of $t - A$. Then by (1), $t - A$ is part of t' .

(1) \Leftarrow (2). Either $t - A$ exists or it does not. Suppose it does, and pick any part s of t that does not overlap a state A is about. We show that s is part of t' . Since $t - A$ is the largest part of t not overlapping $A\text{-part}(t)$, s is part of $A\text{-part}(t)$, and by (2), $A\text{-part}(t)$ is part of t' .

So suppose $t - A$ does not exist. Then by Fact 3.3.3, every part of t overlaps $A\text{-part}(t)$. Then for every part s of t there is a $u \leq s$ with $u \leq A\text{-part}(t) = \sqcup A\text{-parts}(t)$. By no emergent parts, u overlaps some $x \in A\text{-parts}(t)$, i.e. there is some $y \leq u$ such that $y \leq x$ and $x \leq v$ for some state A is about. Then we have $y \leq u \leq s$ and $y \leq x \leq v$, so s overlaps a state A is about. Hence t' is vacuously an A -variant of t . \square

3.6.2 Adding time

The notion of A -variant is, by itself, not enough to capture how we construct hypothetical scenarios, and therefore not enough to account for the meaning of conditionals and causal claims. To see this, consider a light switch connected to a light. Each can be in two states: up/down and on/off, respectively. When the switch is flicked down, the light turns on. For simplicity, suppose the light switch and light are all there is. Our state space is given in Figure 3.20.

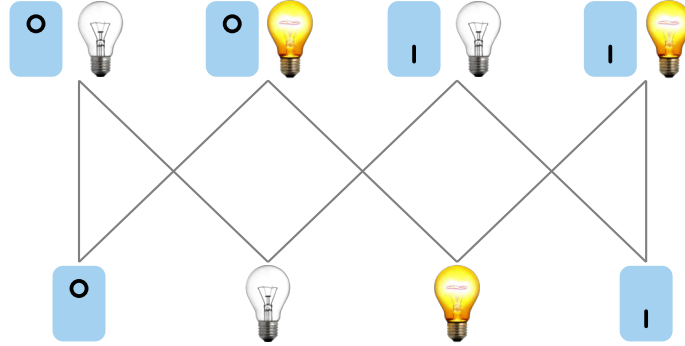
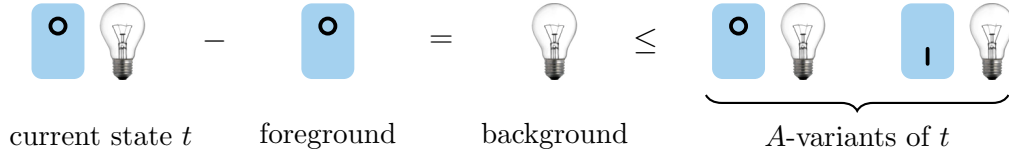


Figure 3.20: A state space of the switch and light.

If we ignored how states can change through time, we might be tempted to say that when asked to imagine A true at world w , the modal horizon is the set of A -variants of w where A is true. This results in a static treatment of sufficiency and the conditional, whereby A is sufficient for C just in case C is true at every A -variant of w where A is true, and the conditional $A > C$ is true just in case C is true at the selected A -variant of w where A is true. Suppose that currently, the switch is up and the light is off is the current state and consider:

- (36) a. If the switch were down, the light would turn on.
 b. The light is off because the switch is up.

These are straightforwardly true. However, the above static treatment of the sufficiency and the conditional predicts them to be false. We assume *the switch is down* is about the states of the switch: $\{\text{○}, \text{I}\}$. Mereologically, the state of the switch does not overlap the state of the light, so when we remove one, the other remains. Then the state of the light being off is part of every A -variant of the current state.



In every A -variant, the light does not turn on. Then if our modal horizon were the A -variants of the current state, we would incorrectly predict (36) to be false.

The larger point is this: mereology and nomic possibility are distinct kinds of structure. The switch and light are nomologically dependent, in the sense that changing the switch leads to a change in the light, but mereologically independent, in the sense that they do not overlap. So in addition to parthood, our model of hypothetical reasoning needs something else. I propose that it needs to specify how the state of the world a given moment in time can change through time. That is, it must specify *nomic possibility*.

We propose to analyse nomic possibility as follows. Recall that we take as primitive a set of states S and a partial order \leq over states, representing parthood. We take each state to be a snapshot, describing how a part of the world stands at a point in time. Since the same state can repeat, let us define a *situation* to be a particular instance of a state.²⁵ Let us then say that a *moment* is a maximal situation with respect to parthood.²⁶ A *world* is linear order of moments – the linear order represents time. This gives us the set of logically possible worlds. From this set we designate a subset P containing all and only the nomically possible worlds. This is summarised in Definition 3.6.3.

²⁵Formally, we take a situation to be a pair (s, i) , often written s_i , where $s \in S$ is a state and i is a label from an arbitrary set of labels. We introduce labels since situations are *particulars* while states are not (i.e. states are multiply realisable, situations are not). In other words, we may think of a situation as a token of a state, following terminology from Tarski and Montague: “Suppose you know all the notions as applied to types. Then as Tarski points out, one should identify a token with a pair consisting of a type and a context” (Montague, 29 August 1967, <https://youtu.be/RkZTF2dilt8?t=225>).

²⁶That is, where s_i is a situation, s_i is a *moment* just in case $s \leq t$ implies $s = t$ for any situation t_j . If one wishes to introduce impossible states – as, say, Fine (2017b, 2021) does – one may designate a set of possible states and define a moment to be a maximally possible state; that is, a moment is a possible state that is not a proper part of any other possible state.

3.6.3. DEFINITION (Imaginative structure). Where S is a set and \leq a binary relation on S , define:

$$\begin{aligned} \text{situations}(S) &:= S \times I, \text{ where } I \text{ is an arbitrary label set} \\ \text{moments}(S, \leq) &:= \{t_i \in \text{situations}(S) : \forall u \in S, t \leq u \Rightarrow t = u\} \\ \text{worlds}(S, \leq) &:= \{(M, \preceq) : M \subseteq \text{moments}(S, \leq), \preceq \text{ is a linear order}\}. \end{aligned}$$

(S, \leq) is a *world-space* just in case it is a partial order and every state $s \in S$ is part of a moment of (S, \leq) .

An *imaginative structure* is a tuple $(S, \leq, P, \mathcal{A})$ where (S, \leq) is a world-space, P a subset of $\text{worlds}(S, \leq)$ and \mathcal{A} is a relation between sentences and states.

To illustrate, in the switch and light example the nomically possible worlds are the directed paths through Figure 3.21 (for simplicity we assume time is discrete and that changes happen after one step in time, though our framework is perfectly capable of representing time as dense). The set of nomic possibilities tells us, for example, that every state where the switch is down is followed by a state where the light is on; it is not nomically possible for the light to spontaneously turn on, without the switch first being flicked down.

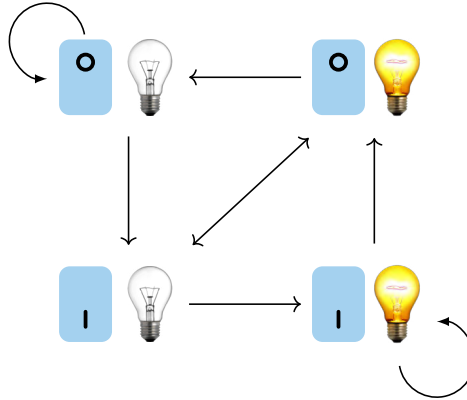


Figure 3.21: The nomically possible worlds correspond to directed paths.

We now have the ingredients to define the modal horizon.

(37) **Proposal.** $\text{int}(t, A)$ is the set of A -variants of t .

We then define what it means to ‘play the laws’ forward from a moment.

(38) For any world w , sentence A , set of worlds P and moment t , define

$$\text{mh}_{P,t}(w, A) := \{w_{\prec t} \frown w'_{\succeq t'} : t' \in \text{int}(t, A), t' \in w' \text{ and } w' \in P\}$$

Our definition of interventions is illustrated in Figure 3.22.

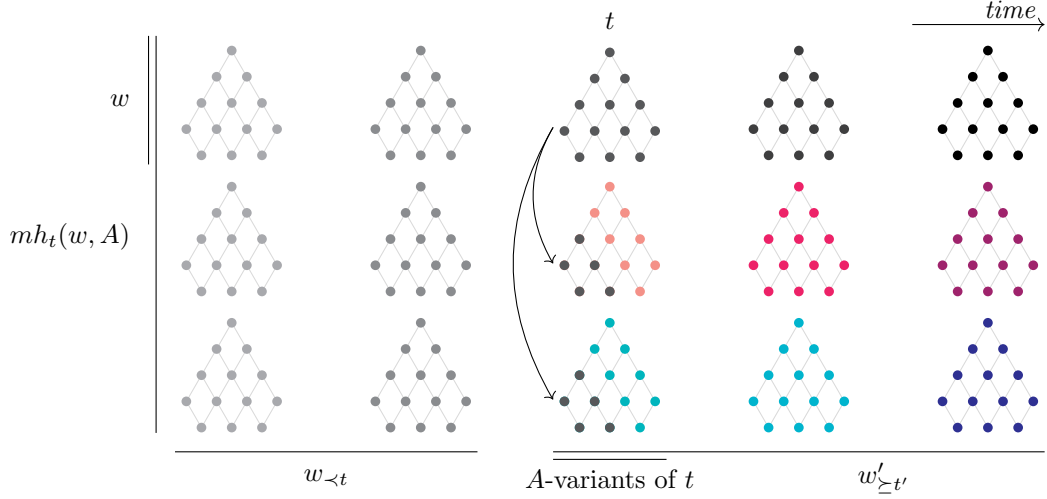


Figure 3.22: How to construct the modal horizon.

Lastly, recall our proposal in section 3.2.4 whereby A is sufficient for C ($A \gg C$) at a world w just in case C is true at every A -world in the modal horizon $mh_{P,t}(w, A)$, and the conditional $A > C$ is true at w just in case the world C is true at the selected A -worlds in modal horizon $mh_{P,t}(w, A)$.

- (18) Where P is the set of nomically possible worlds, t the intervention time, and s the selection function,

$$\begin{aligned} A \gg C \text{ is true at } w & \quad \text{iff} \quad mh_{P,t}(w, A) \cap |A| \subseteq |C| \\ A > C \text{ is true at } w & \quad \text{iff} \quad s(w, mh_{P,t}(w, A) \cap |A|) \in |C| \end{aligned}$$

With these entries we correctly predict the truth of (36), repeated below.

- (36) a. If the switch were down, the light would turn on.
b. The light is off because the switch is up.

Currently, the switch is down and the light is off. Where A is the sentence *the switch is down/not up*, recall that A -variants of the current state are the moments containing the state of the light being off. When we then restrict to those A -variants where A is true (i.e. where the switch is down), we see that in all possible continuations of this moment, the light turns on. So at this moment the switch being down is sufficient for the light to be on and predict (36a) to be true. Moreover, when we restrict to the A -variants where the switch is up, every possible continuation contains a state where the light is off, so the switch being

up is sufficient for the light to be off. Since both of these sufficiencies hold, we predict (36b) to be true, as desired.

3.6.3 Imagining changes to the laws

In addition to imagining changes to ordinary properties like shape and colour, we can imagine changes to the laws of nature themselves. For example, it is often said that the physical constants are finely tuned for stars to form (Smolin 2013). In *A Brief History of Time*, Hawking writes

- (39) If the electric charge of the electron had been only slightly different, stars either would have been unable to burn hydrogen and helium, or else they would not have exploded. (Hawking 1988)

As (39) shows, we use the same linguistic construction to consider changes to the laws as we do changes to ordinary properties. This suggests that there is no difference in principle between imagining a change to ordinary properties or the laws: they make use of the same general imaginative faculty. We would therefore like a single framework that can represent both.

The way we imagine changes to the laws is often structured, or systematic, in the sense that we can imagine changing some laws while fixing other facts. We can imagine what would happen if the gravitational constant were different but the fine-structure constant the same. Or imagine a mug filled with hot coffee and consider (40).

- (40) If the mug suddenly turned into ice, the coffee would melt it.

This has a true reading. Now, we may suppose it is not nomically possible for mugs to suddenly turn to ice. Nonetheless, we can imagine this change while fixing the other features of the scenario, such as the fact that there is hot coffee in the mug, as well as the other laws, such as the fact that heat melts ice.

How can we represent structured changes to the laws in the present framework? Note that on this approach the state space and nomic possibility have a different formal character: the state space is *structured* while nomic possibility is *unstructured*. A world is either nomically possible or not – that’s all there is to the concept. The present approach uses the structure of the state space to account for the systematic nature of hypothetical reasoning: the fact that we can imagine changing some states while fixing others. One might therefore think that the present framework cannot represent the systematic way in which we imagine changes to the laws, since doing so would cast us out of the realm of nomic possibility altogether and into an unstructured wilderness.

In fact, the present approach can represent the systematic way in which we can imagine changes to the laws. The idea is that when we imagine changes to the laws, we are not imagining a change to what worlds are nomically possible,

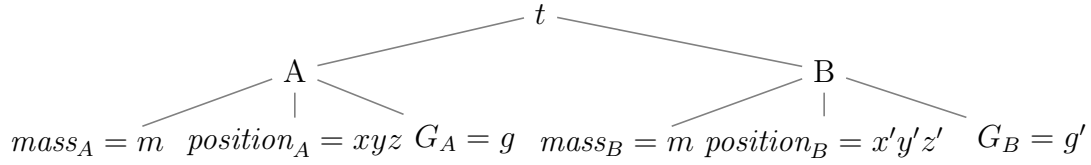
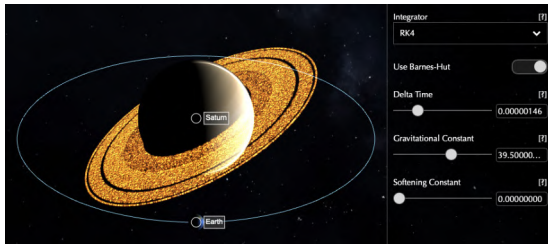


Figure 3.24

but simply moving to a different set of nomically possible worlds. To do this, we represent changes to the laws as changes to an object's properties, and let the set of nomic possibilities include states where objects have these different properties.²⁷ In section 3.4 we proposed that we can represent an object's properties as parts of it, such as its colour and material. Another property of objects we may consider is how things interact with their environment. By including states representing how an object behaves, we can represent changing the laws in a systematic way. For just as we can change an object's colour or shape while keeping its other properties, we can change one property of how an object interacts with its environment while keeping the others.

Figure 3.23: From gravitysimulator.org.

For example, we may represent the value of the gravitational constant as a property of a scenario, as in a physics simulator such as that in Figure 3.23, featuring a slider that allows the user to change the gravitational constant.

Alternatively, we may represent different objects having a different gravitational constant (that is, different objects with the same mass attracting other objects with different force). In such cases our concept of nomic possibility is still defined; for example, we can imagine what would happen if bodies A and B had the same mass, but A nevertheless attracted objects with a greater gravitational force than B. In that case, if A and B begin at a distance, then they will meet closer to A's original position than to B's (assuming no external forces). We can represent this reasoning on the present approach by taking, say, the value of an object's gravitational constant as a property of it, as in Figure 3.24. Since the states $G_A = g$ and $G_B = g'$ do not overlap in this model, the present approach predicts that we can vary the gravitational constant of one while fixing that of the other. To capture our ability to reason with objects of with different gravitational constants, we refine the nomic possibilities accordingly. For example, we might ordinarily say that in every Newtonian nomically possible world containing two bodies of equal mass, the bodies will meet at the

²⁷For discussion of the idea that nomic possibility is determined the properties of objects, see Fine (2002) and Lowe (2005), among others.

midpoint of their centres of mass (assuming no external forces). This implicitly assumes that all bodies have the same gravitational constant. If we drop that assumption, a moment such as t in Figure 3.24, where $g \neq g'$, is still in the realm of nomic possibility, so our framework can make predictions about what would happen in the future of such a moment, such as where the two bodies would meet.

3.6.4 The intervention time

Our definition of the modal horizon itself does not place any constraints on which intervention time we pick. Some choices are ruled out by how the meaning of conditionals and causal claims interact with general principles of conversation. Let us say that an intervention time t is *trivial* with respect to w and sentence A just in case A is true at every world containing $w_{\prec t}$ or false at every world containing $w_{\prec t}$, where $w_{\prec t}$ the initial segment of w up to t . Principles of conversation require choosing a non-trivial intervention time. For example, when a conditional antecedent is actually false, picking a trivial intervention time results in a trivial meaning for the conditional. For then, given that interventions fix the world up to intervention time, A is false at every world in the modal horizon, $mh_{P,t}(w, A) \cap |A|$ is empty, so $A > C$ is either vacuously true or suffers from presupposition failure (depending on one's preferred diagnosis). Either way the conditional is uninformative, ruling out an intervention time after the antecedent becomes false. And when A is actually true, picking a trivial intervention time violates Condoravdi's (2002) diversity condition; for then A is true throughout $mh_{P,t}(w, A)$, so $A > C$ is equivalent to $A \wedge C$ (by the centering requirement of the selection function), making the conditional construction – and the complex semantics it brings with it – redundant.²⁸

Furthermore, setting the intervention time of a *cause* or *because* claim to after the cause argument is settled results in a trivially false claim. If the cause argument is false this is simply because causal claims entail that their cause occurred. If the cause argument is true, a *cause* or *because* claim is trivially false because it entails that the cause is not sufficient for it to produce the effect, which in turn entails that there is a world in the modal horizon where the cause argument is false; in symbols, $\neg(\neg C \gg (\neg C \text{ produce } E))$ entails that $mh_{P,t}(w, \neg C) \cap |\neg C|$ is nonempty. But the modal horizon cannot contain a $\neg C$ -world if we pick an intervention time after C has already become true.

While principles of conversation rule out some intervention times, they do not determine the intervention time uniquely. To introduce the effect of different intervention times, consider the image in Figure 3.25 by James Fridman, a graphic designer known for his surprising photo edits.

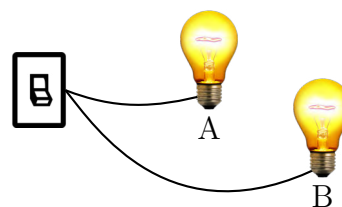
²⁸A number of authors also propose that tense in conditionals can shift when hypothetical possibilities are accessed (Arregui 2007, Ippolito 2013, Khoo 2015, 2017). On these accounts, tense in conditionals also constrains what we call the intervention time.



Figure 3.25: One of James Fridman’s [photo edits](#). Used with permission.

At the risk of analysing the humour away, Fridman’s joke is a play on intervention time. The person requesting the edit is hoping for a sudden intervention, one where the wet floor sign miraculously vanishes and we see the resulting image exactly at intervention time. In Fridman’s reply, the intervention time and the time when we evaluate the image are split. The sign disappears, and some time passes before we see the image.

Let us turn to the role of intervention time in our interpretation of conditionals. Consider a common cause structure, such as a light switch connected to lights A and B, depicted on the right.²⁹ When the switch is flicked down, light A turns on, and a bit later light B turns on. Suppose that both lights were off. The switch was flicked down. Light A turned on, then light B turned on. Consider (41) in this context.



- (41) a. If light A were off, light B would be off.
 b. If light A turned off, light B would turn off.

Intuitively, these have a true reading in this scenario, called a *backtracking reading* (see e.g. Lewis 1979, Arregui 2005, Khoo 2017).³⁰ This suggests that when

²⁹For previous discussion of the semantics of conditionals in common cause scenarios, see Hiddleston (2005).

³⁰We use both *were off* and *turned off* to show that the true reading is available for both stative (*is off*) and eventive (*turn off*) antecedents.

we interpret (41) we can vary the position of the switch. We can show this more directly with (42).

- (42) a. If light A were off, the switch would be in a different position.
 b. If light A turned off, the switch would be in a different position.

These also have a true reading.³¹ This shows that the intervention time does not have to be the time when the eventuality described by the conditional antecedent occurs. For example, when we interpret (41b) and (42b) we intervene before light A turns off.

We can account for the acceptability of the sentences in (41) and (42) by assuming that the state of the switch being down is exactly relevant to *Light A is off*. For given that this state is also nomically relevant to *Light A is off*, when we interpret these sentences we allow the state of the switch to vary. When we restrict to the nomically possible worlds containing the remaining states where light A is off, we find light B is also off, so we predict (41) and (42) to be true.

Turning to causal claims, consider the following sentences in the same context.

- (43) a. Light A turning on caused light B to turn on.
 b. Light B turned on because light A turned on.

These sentences are intuitively unacceptable. Among counterfactual dependence approaches to causation, a typical explanation for their falsity is that causal claims do not allow for backtracking interpretations (see e.g. Lewis 1979). The idea is that when we intervene to turn light A off on a non-backtracking reading, the switch does not change, so light B still does not turn off.

If *C cause E* and *E because C* required there to be some world in $mh_{P,t}(w, \neg C)$ where *E* is false, a ban on backtracking would indeed account for the unacceptability of (43). However, we have independent evidence that *cause* and *because*

³¹Lewis (1979:458) writes that “Back-tracking counterfactuals, used in a context that favors their truth, are marked by a syntactic peculiarity. They are the ones in which the usual subjunctive conditional constructions are readily replaced by more complicated constructions: ‘If it were that... then it would have to be that...’ or the like.” (for discussion see Arregui 2005:Ch. 3). The acceptability of (41) and (42) shows that backtracking readings do not require this ‘syntactic peculiarity’. There is, nonetheless, a contrast between the following.

- (i) a. ??If light A were off, the switch would move to a different position.
 b. ??If light A turned off, the switch would move to different position.
 (ii) a. If light A were off, the switch would have to have moved to a different position.
 b. If light A turned off, the switch would have to have moved to different position.

Enç (1996) and Condoravdi (2002) propose that modals shift the time of evaluation of their prejacent. Following this idea, we propose that the time when a conditional consequent is evaluated begins with the time when the antecedent is true, rather than the intervention time. This accounts for the unacceptability of (i). The extra *have* in (ii) backshifts the consequent’s evaluation time, rescuing them.

claims can be true even when the effect would have happened anyway (e.g. the Billy and Suzy case from Hall 2004). Such cases of overdetermination show that *C cause E* and *E because C* can be true even when *E* is true at every world in $mh_{P,t}(w, \neg C)$.

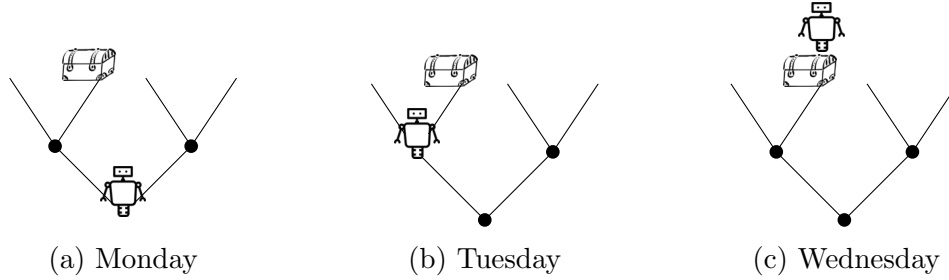
Following Beckers (2016), and put in terms of the present account, we propose that *C cause E* and *E because C* require the truth of $\neg(\neg C \gg (\neg C \text{ produce } E))$, i.e. they require the existence of a world in $mh_{P,t}(w, \neg C)$ where $\neg C$ does not produce *E*. According to the analysis of production we offer, in the common-cause context above every world in $mh_{P,t}(w, \neg C)$ is one where light A turning off does not produce light B to turn on. A fortiori, there is a world in $mh_{P,t}(w, \text{Light A turns off})$ where light A turning off does not produce light B to turn on. But far from predicting the causal claims to be unacceptable, this is precisely what *cause* and *because* require. We therefore need an alternative account of why (43) are unacceptable. We propose they are unacceptable since *cause* and *because* entail that the cause produced the effect (see chapter 5 for an analysis of production).

As well as accounting for overdetermination readings, this account preserves the idea from section 3.2 that counterfactuals and causal claims raise hypothetical scenarios in a uniform way. For we can maintain that the modal horizon $mh_{P,t}(w, A)$ is present in the semantics of both conditionals and causal claims. We do not need to stipulate that causal claims use a special non-backtracking modal horizon. Rather, we trace the difference between (41) and (43) to an independently-motivated difference between conditionals and causatives: the fact that causatives involve production while conditionals do not.

Now, if our uniformity hypothesis – that conditionals and causal claims raise hypothetical scenarios in a uniform way – is correct, and the intervention time can backshift when interpreting conditionals, we would also expect it to be able to backshift when we interpret causal claims. In other words, just as conditionals have backtracking readings, we would expect causal claims to have backtracking readings too. This is a surprising, perhaps radical prediction of our uniformity hypothesis, in light of the longstanding stipulation among counterfactual dependence analyses of causation that backtracking readings are forbidden in the interpretation of causal claims.

Remarkably, this prediction appears to be borne out. Recall the robot context. Suppose the robot has to take walk, with treasure at the end of one of the roads. When the robot reaches a fork in the road, it decides randomly which way to turn. On Monday morning the robot was placed on the starting point. It then turned left on Monday at random, and on Tuesday turned right at random, reaching the treasure. Consider (44) in this context.

- (44) a. The robot reached the treasure because it turned right on Tuesday.
 b. The robot turning right on Tuesday caused it to reach the treasure.



(44) can sound quite good. They also appear to have a reading on which they are not acceptable. This reading is subtle, but can be brought out by considering that the robot could have instead turned right on Monday and then again right on Tuesday. In that case it would have missed the treasure. We make this reading of (44) salient in the sentences below, which while cumbersome, are nonetheless interpretable.

- (45) a. Given how things were on Monday morning, the robot reached the treasure because it turned right on Tuesday.
 b. Given how things were on Monday morning, the fact that the robot turned right on Tuesday caused it to reach the treasure.

The fact that (44) have a false reading shows, I believe, that when we interpret (44) we do not have to fix the fact that the robot turned left on Monday. This suggests that when we interpret a causal claim we do not have to fix things that happened before the time when the cause became true.³² In other words, just as there are backtracking conditionals, there are also *backtracking causatives*.³³

The existence of backtracking readings of causatives is a problem for theories that account for the unacceptability of (43) by banning backtracking in the interpretation of *cause* and *because*. Such theories have to explain why backtracking is forbidden in (41) but not in (44). In contrast, backtracking causatives are not a problem on our account since we never needed to ban backtracking. We instead accounted for the unacceptability of (43) using an independently-motivated difference between causatives and conditionals: the fact that causatives, but not conditionals, involve production. For given that *C cause E* and *E because C*

³²Though we seem to have a default preference for later intervention times (see Khoo 2017).

³³Shifts in intervention time seem especially easy with stative causes. For example, suppose Ali's parents are monolingual Farsi speakers and raised him speaking Farsi, and consider:

- (i) a. Ali understands what his parents are saying because he speaks Farsi.
 b. The fact that Ali speaks Farsi causes him to understand what his parents are saying.

These sentences have a true reading. If, however, we emphasise that Ali's parents taught him to speak, we can reason that if Ali didn't speak Farsi he would still speak his parents' language, and so understand what they are saying. On this reading the sentences in (i) are unacceptable.

require that C produce E , but light A turning on did not produce light B to turn on, we predict (43) to be unacceptable. The upshot is that we get to keep the idea that conditionals and causatives raise hypothetical scenarios in a uniform way; that we have a general ability to consider hypothetical scenarios in response to a sentence which we use when we interpret conditionals and causatives alike.

3.7 Exploring the present proposal

The proposal in (18) leaves many things open; namely, what the aboutness relation is, when the intervention time should be, which worlds are nomically possible, and which selection function is used. Nonetheless, our proposal already makes some concrete predictions about sufficiency and *would*-conditionals.

We predict the strength ordering in Figure 3.27: if A entails C then A is sufficient for C , which in turn implies the truth of the conditional $A > C$, which in turn implies the truth of the material conditional $A \supset C$, as expected.

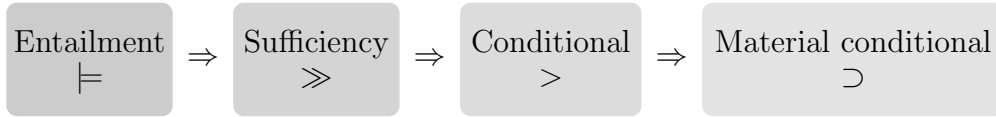


Figure 3.27: Strength ordering between four relations of inference.

Entailment implies sufficiency since, given that every world where A is true, C is true, then in particular, every world in the modal horizon where A is true, C is true. Sufficiency implies the conditional because s is a selection function: by the success condition in (14), the world $s(w, p)$ is selected from among the worlds where p true; so if C holds throughout the set $mh_{P,t}(w, A) \cap |A|$, then C also holds at the world selected from this set. Lastly, the conditional implies the material conditional due to the selection function and the fact that our proposal validates the following analogue of weak centering: every nomically possible world is in its own modal horizon.

- (46) For every sentence A , moment t , set of worlds P and world $w \in P$,
 $w \in mh_{P,t}(w, A)$.

To see why this holds, note that every moment is an A -variant of itself. This follows immediately from the definition of A -variant: for any moment t and sentence A , every part of t that does not overlap a state A is about is, of course, part of t . Then as the actual world w is nomically possible, w (which is just the concatenation of its past, present and future, $w = w_{<t} \frown w_{\geq t}$) is in the modal horizon $mh_{P,t}(w, A)$. And because the selection function is strongly centered, if A is true at w then w is in $mh_{P,t}(w, A)$, so the selection function must select it. Altogether, then, if $A > C$ and A are true at w , C is too.

Since consequents are evaluated using ordinary truth at a world, our proposal validates the familiar rules of triviality ($A \gg \top$; $A > \top$), identity ($A \gg A$; $A > A$), right weakening,

$$\frac{A \gg B \quad B \models C}{A \gg C} \qquad \frac{A > B \quad B \models C}{A > C}$$

deduction within conditionals

$$\frac{B_1 \wedge B_2 \wedge \dots \models C}{(A \gg B_1) \wedge (A \gg B_2) \wedge \dots \models A \gg C} \qquad \frac{B_1 \wedge B_2 \wedge \dots \models C}{(A > B_1) \wedge (A > B_2) \wedge \dots \models A > C}$$

and the conjunction rule

$$\frac{A \gg C_1 \quad A \gg C_2 \quad \dots}{A \gg (C_1 \wedge C_2 \wedge \dots)} \qquad \frac{A > C_1 \quad A > C_2 \quad \dots}{A > (C_1 \wedge C_2 \wedge \dots)}.$$

Infamously, Lewis's (1973) semantics does not validate the conjunction rule for conditionals, owing to failures of the limit assumption. We will further discuss this rule in section 4.3, when we consider scenarios where our intuitive concept of similarity violates the limit assumption. In contrast, the present approach validates the conjunction rules 'out of the box' – without terms and conditions.

Note that on the present approach the rules discussed in this section are valid regardless of how we fill in the parameters of interpretation; that is, regardless how we analyse the notion of aboutness, what intervention time we choose, which worlds are nomically possible and which selection function is at play. These validities are guaranteed by the formal architecture of the account alone. This is a welcome result, since these principles are some of the most incontrovertible facts we have about sufficiency and the conditional.

3.7.1 What is aboutness?

The foreground of a sentence is the set of states we allow to vary when that sentence appears as a conditional antecedent or as the cause argument in a causal claim. We took a sentence's foreground to be the set of states it is in some sense about. What does it mean for a sentence to be 'about' a state? A natural idea is that a sentence is about the states that are in some sense 'directly responsible' for the sentence having the truth value it has. We will explore one way to analyse this idea, by decomposing it into two relevance conditions. First, a *nomie relevance* condition: every state sentence A is about is 'nomically relevant' to A , in a sense to be made precise; second, an *exact relevance* condition: every state A is about is in some sense exactly relevant to the truth or falsity of A .

- (47) **Proposal.** A sentence A is about a state s just in case s is nomically relevant to A and exactly relevant to A .

We analyse each notion in turn.

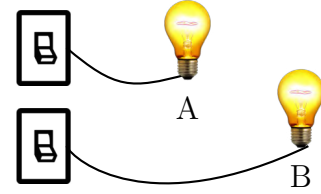
Nomic relevance. Let us define nomic relevance as follows.³⁴

- (48) State s is nomically relevant to sentence A just in case
- a. A is true at every nomically possible world containing s ; or
 - b. A is false at every nomically world containing s ; or
 - c. A is true at every nomically possible world not containing s ; or
 - d. A is false at every nomically possible world not containing s .

Note that our definition of nomic relevance takes into account which worlds are nomically possible. So we predict that what a sentence is about also depends on the nomic possibilities.

To test this prediction, imagine a variant of the common cause structure in section 3.6.4, where this time each light has its own switch. Consider (41) again.

- (41) a. If light A were off, light B would be off.
b. If light A turned off, light B would turn off.



In this context (41) are unacceptable. Recall that they were fine in the one-switch context. Why the contrast?

In the one-switch context, the state of the switch connected to light B is nomically relevant to *Light A is off*, while in the two-switch context it is not (Figure 3.28).

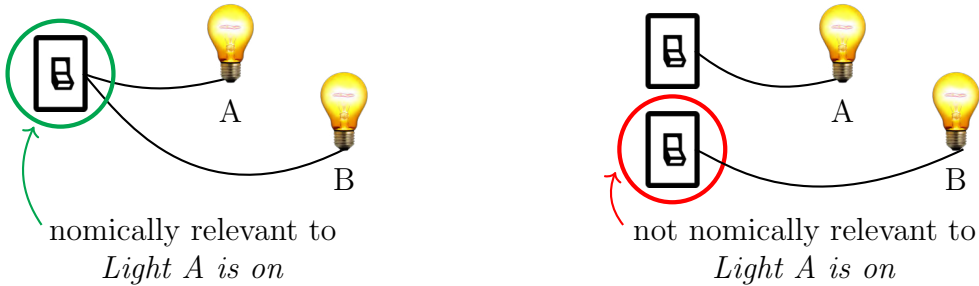


Figure 3.28

The nomic relevance constraint ensures that in the two-switch context *Light A is off* is not about the state of B's switch, so when we interpret (41) in the two-switch context we fix the state of B's switch. For every moment containing the state of switch B being up, every nomically possible future of this moment

³⁴This definition is inspired by the definition of orthogonality from Lewis (1988). Formally, a world $w = (M, \preceq)$ contains a state s just in case $s \leq t$ for some moment $t_i \in M$.

is one where light B is on. So given the nomic relevance constraint, we correctly predict (41) to be false in the two-switch context.

The nomic relevance constraint helps account for a wide range of facts we intuitively fix when we construct hypothetical scenarios. Take (49).

- (49) a. If I had pushed the power button, the TV would have turned on.
 b. *Context: I pushed the power button and the TV turned on.*
 (i) Pushing the power button caused the TV to turn on.
 (ii) The TV turned on because I pushed the power button.

We can accept these sentences even though pushing the button is not sufficient by itself for the TV to turn on. When we evaluate them we fix the fact that the TV is plugged in, that the electricity is working and countless other states, too many to explicitly list. These are not nomically relevant to *I pushed the power button*. For example, there is a world where the electricity is working and I push the button, a world with electricity where I don't, a world without electricity where I do, and a world without electricity where I don't. The nomic relevance constraint therefore correctly predicts that the sentence *I pushed the power button* is not about these states.

The nomic relevance constraint also helps account for the fact that when we construct hypothetical alternatives, we do not imagine changes to the laws themselves unless explicitly told to do so (an observation made by Seelau et al. 1995, Byrne 2005:10 and Veltman 2005, among others). Consider (50).

- (50) If Rosie had let go of the ball, it would have fallen.

As a matter of fact, when we interpret (50) we fix the fact that there is gravity. How do we intuitively know to do so? We may suppose that part of the scenario, call this state GRAVITY, determines that the ball falls when dropped. To capture the fact that we fix GRAVITY when we interpret (50), we need $A = \text{Rosie let go of the ball}$ to be about this state. Happily, it turns out that the nomic relevance constraint guarantees this. Given our definition of nomic relevance in (48), GRAVITY is not nomically relevant to A . This is because there are worlds with gravity where Rosie lets go of the ball, worlds with gravity where she holds on to it, worlds without gravity where she lets it go, and worlds without gravity where she holds on to it. Compare this with (51).

- (51) If Rosie had let go of the ball and gravity had stopped working, it would have fallen.


While GRAVITY is not nomically relevant to A , it is nomically relevant to $A \wedge B = \text{Rosie let go of the ball and gravity stops working}$, since every world containing GRAVITY is a world where $A \wedge B$ is false. This allows $A \wedge B$ to be about the state GRAVITY, as desired.

In general, then, for any sentence A that does not mention the laws, the parts

of the world that guarantee the truth of the laws are not nomically relevant to *A*. The nomic relevance constraint therefore predicts that when we imagine *A* true, we do not imagine changes to the laws.

Exactness. By itself, nomic relevance is too permissive in what states it allows a sentence to be about. The way we construct hypothetical scenarios is often surgical, varying a part of the world at a moment in time while leaving the rest intact. We saw this in our initial example, repeated below.


- (1) a. The light is off because switch A is down.
- b. If switch A were up, the light would be on.

When we interpret (1) we vary the position of switch A but not switch B. Now, the state of both switches  is nomically relevant to *switch A is down* but this sentence is not about this state. Intuitively, *switch A is down* is not about the state of both switches because that state has a part that is irrelevant to the truth of sentence: the state of switch B. We therefore add an ‘exact relevance’ constraint:

- (52) For every sentence *A*, every state *A* is about is exactly relevant to *A*.

Since a state can be nomically relevant to a sentence without the sentence being about the state, it is the exactness of exact relevance that accounts for the exactness of our imagination when we consider hypothetical scenarios in response to a sentence.

The big question is what it means for a state to be exactly relevant to a sentence. Philosophers have recently paid much attention to this question, particularly in work on aboutness (Yablo 2014, Hawke 2018, Berto 2018) and truthmaker semantics (Fine 2017b). Fine (2017b) introduces two relations, called *exact verification* and *exact falsification*, between states and sentences. The guiding idea is that a state exactly verifies a sentence just in case it is wholly relevant to the truth of the sentence, and exactly falsifies a sentence just in case it is wholly relevant to the falsity of the sentence.

While Fine does not analyse what it means for a state to exactly verify or falsify a sentence, in many cases the notion is perfectly clear. For example, the state  exactly verifies *switch A is down* and exactly falsifies *switch A is up*, and it is the only state of Figure 3.18 to do so.

An longstanding issue in the literature on exact relevance is the relationship between exactness and minimality. Let us say that a state *settles* a sentence just in case the sentence is true at every nomically possible world containing the state or false at every nomically possible world containing the state. We also say that a state *minimally settles* a sentence just in case it settles the sentence and no proper part of the state also settles the sentence. Can a sentence be about a

state without the state without minimally settling the sentence? To answer this, consider (53).

- (53) a. There is mud.
 b. There are infinitely many stars.
 c. This is moving.

Kratzer (2012:166) points out that, if we assume mud is mud ‘all the way down’ – every proper part of mud is itself mud – then the sentence *There is mud* has no minimal situations where it is true. Neither does *There are infinitely many stars* (Kratzer 1990, 2002, 2012:171). Every state containing infinitely many stars has a proper part (a smaller infinite collection of stars) that still settles the sentence, and a proper part (a finite collection of stars) that does not settle the sentence. Fine also warns against defining exactness in terms of minimality, pointing out that for the sentence *This is moving*, “we may well maintain that any verifier (the motion of the object through an interval of time) will contain another verifier as a proper part” (2017:564).

Now consider what scenarios we imagine when the sentences in (53) appear as a *because*-clause or conditional antecedent. For example, suppose there are infinitely many stars. Then no state minimally settles that there are infinitely many stars. So if minimally settling a sentence were required for the sentence to be about the state, we would predict that every state is held fixed when we interpret the antecedent $A = \textit{If there were finitely many stars, ...}$. But then we would keep the fact that there are infinitely many stars and would predict A to be unimaginable, in the sense that there is no world in A ’s modal horizon where A is true. Clearly, *there are infinitely many stars* must be about a state consisting of infinitely many stars even though that state does not minimally settle the sentence.

The lesson I draw from these examples is that aboutness depends not only on which states settle a sentence but also on its logical structure.³⁵ Notice that the sentences in (53) are all in some sense logically complex. For example, *There is mud* is a quantified sentence. A logically simpler sentence, such as *Region r is mud*, has a minimal state settling it: r itself – assuming ‘is mud’ means ‘is mud throughout’. For any proper subregion r' of r does not verify that region r is mud throughout. One can proceed to analyse other expressions in terms of this predicate and logical terms; for example, we can analyse *Region r contains mud* as *r has a subregion that is mud*.

Or take *This is moving*. In logic class – say, when introducing propositional or first-order logic – we would typically call this an atomic sentence. At the same time, one can decompose its meaning in terms of tense and aspect. Adopting event semantics (e.g. Davidson 1967a, Parsons 1990), for example, and adopting

³⁵Other approaches that appeal to logical structure include Fine’s truthmaker semantics (Fine 2017b) and the proposal by Deigan (2020).

a standard meaning for the imperfective (Klein 1994, Kratzer 1998), we can represent the meaning of *This is moving* as

$$\exists e(\text{move}(e) \wedge \text{agent}(e) = x \wedge \text{runtime}(e) \subseteq t),$$

where e ranges over eventualities (events or states), x is the referent of *this* and t the time of evaluation.³⁶ Just as we assumed ‘is mud’ means ‘is mud throughout’, we assume that *move* in the formula above means ‘move throughout’. Then e minimally settles *move*(e). In general, we can take our atomic sentences to be of the form $P(e)$, where e is an eventuality.³⁷

The sentences in (53), then, are not obstacles to defining exact relevance in terms of minimality plus logical structure; that is, defining exact relevance in terms of minimality for atomic sentences and extending this compositionally to logically complex sentences with exact relevance clauses for the logical terms. One might therefore be tempted to say that an atomic sentence is about the set of states that minimally settle it. (This proposal automatically satisfies the nomic relevance constraint.) As a matter of fact, for every atomic sentence we consider in this essay, the states we assume it is about are indeed the states that minimally settle it. Nonetheless, there is clearly much more to say on this topic than we can here, so we will adopt a tolerant approach: nothing we say in what follows will assume that exact relevance for atomic sentences can be defined in terms of minimality. One does not need to assume this to apply the present framework. Chapter 4 we discuss ways to define exact relevance for logically complex sentences in terms of exact relevance for atomic sentences, which does not require defining what it means for a state to be exactly relevant to an atomic sentence.

3.7.2 Aboutness for random variables

In section 3.4.3 we saw how to generate a state space from a set of random variables, such as those used in structural causal models (Pearl 2000). Random

³⁶It is natural to wonder what part of the world is responsible for the truth of sentences such as $\text{agent}(e) = x$ and $\text{runtime}(e) \subseteq t$. We address this in section 4.2.4.

³⁷Though one may decompose further, classifying sentences of the form $P(e)$ as non-atomic. For example in lexical semantics one can decompose *open*(e) (event e is an opening event) using operators such as ACT, CAUSE and BECOME (Dowty 1979, Grimshaw 1993). Some commonly cited evidence for this view comes from scope ambiguities with adverbs such as *again* (e.g. Morgan 1969, McCawley 1973, Dowty 1979, von Stechow 1995), as in (i) from Levin (2005).

- (i) Tracy opened the door again.
- a. Repetitive reading: [**again** [[Tracy ACT] CAUSE [BECOME [door OPEN]]]]
‘Tracy yet again performed the activity of opening the door.’
 - b. Restitutive reading: [[Tracy ACT] CAUSE [BECOME [**again** [door OPEN]]]]
‘Tracy brought it about that the door was once more open
(though she may not have opened the door previously).’

variables also generate a natural interpretation of aboutness. For any variable X with value x , $X = x$ be about the set of variable assignments that assign a value to X and only to X . Similarly, where \vec{X} is a sequence of variables and \vec{x} a sequences of values for each variable, let $\vec{X} = \vec{x}$ be about the variable assignments that assign a value all and only the variables in \vec{X} .³⁸

$$\vec{X} = \vec{x} \text{ is about } s \quad \text{just in case} \quad s : \vec{X} \rightarrow R(\vec{X})$$

In work applying structural causal models to conditionals and causal claims it is common to restrict causes and conditional antecedents to conjunctions of literals.³⁹ Our procedure for generating state spaces and aboutness from random variables shows that the present approach matches the expressive power of approaches to conditionals and causal claims that use random variables. Going further, chapter 4 considers ways to extend the aboutness for atomic sentences to aboutness for logically complex sentences.

3.7.3 Capturing sufficiency violations

We can distinguish two ways for sufficiency to fail: due to the openness of the future, and due to the various ways for sentences to be true. The robot example in (2) show the first kind, while the passport examples in (3) show the second.

- (2) *The robot turns at random. It took First Street and then Road B.*
 - a. The robot took Road B because it took First Street.
 - b. The robot taking First Street caused it to take Road B.
- (3)
 - a. Ali has an Irish passport because he was born in Europe.
 - b. The fact that Ali was born in Europe caused him to get an Irish passport.

Our proposal can capture both kinds of sufficiency violation. This is because our definition of the modal horizon takes into account both A -variants, giving us the different ways for the cause to hold, as well as the nomically possible continuations of each A -variant, giving us the openness of the future.

Take the robot example. We assume the intervention time t is just before the robot took First Street. Intuitively, there is only one way for the robot to take

³⁸An alternative approach, one similar to a proposal by Fine (2012b:243), is to let $\vec{X} = \vec{x}$ be about the *fusions* of the variable assignments that assign a value all and only the variables in \vec{X} . This is equivalent to the approach adopted above given our use of overlap in the definition of A -variants (Definition 3.6.1), since for any set of states T with fusion $\bigsqcup T$, every state in T overlaps $\bigsqcup T$, so when we remove one, we remove the other too.

³⁹For example, Pearl (2000:§7.4.2) restricts his semantics of counterfactuals to conjunctions of literals, and Halpern (2016) and Beckers and Vennekens (2018) restrict causes to conjunctions of literals. Though see Briggs (2012) and Ciardelli, Zhang, and Champollion (2018) for semantics of conditionals based on structural causal models that allow for sentences of greater logical complexity.

First Street. That is, there is only one *The robot took First Street*-variant of t where the robot takes First Street; namely, t itself. But there are two nomically possible continuations of t : one where it takes Road A and one where it takes Road B. So among the worlds in the modal horizon where the robot takes First Street is one where it takes Road A. That is, we predict that the robot taking First Street was not sufficient for it to take Road A, and since *cause* and *because* imply that the cause was sufficient for the effect, we correctly predict (2) to be unacceptable.

Turning to the passport case, we suppose that the event of Ali's birth is exactly relevant to *Ali was born in Europe*. Given that Ali was born in Ireland, intuitively, every part of the world in which he was born in Europe is a part of the world in which he was born in Ireland. There is no indeterminate state of Ali being born somewhere or other in Europe. Then every state that is part of the actual world in which *Ali was born in Europe* is about is – or at least overlaps – a state in which Ali was born in Ireland, and is therefore removed when we interpret (3). The actual world has *Ali was born in Europe*-variants where Ali was still born in Europe but outside Ireland. Given that, in some of these worlds, Ali does not have an Irish passport, *Ali was born in Europe* is not sufficient for *Ali has an Irish passport*, and we predict (3) to be unacceptable, as desired.

For a second example, one with a simpler state space, recall (9).

- (9) *Let x and y be numbers, where $x \neq 0$ and $y = 0$.*
 a. *xy is 0 because y is less than 10.*
 b. *xy is 0 because y is 0.*

Consider a state space where each state represents an assignment of values to some variables, depicted in Figure 3.29.

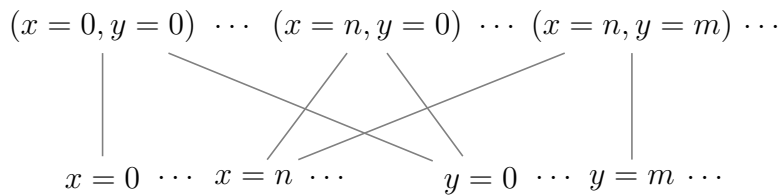


Figure 3.29

We are given in (9) that $x \neq 0$, and so $x = n$, for some nonzero n . Then the actual world of this toy state space is $(x = n, y = 0)$. We assume that $y = 0$ is exactly relevant to *y is 0* and *y is less than 10*. There is no indeterminate state settling that y is some or other number less than 10 without settling what precise number y is. Since the state $x = n$ does not overlap the state $y = 0$, a world is a *y is less than 10*-variant of $(x = n, y = 0)$ just in case it contains the state $x = n$. When we restrict to those variants where y is less than 10, we find plenty where

xy is not zero: $(x = n, y = 1)$, $(x = n, y = 2)$, and so on. Then y is less than 10 is not sufficient for xy is 0 and we predict (9a) to be unacceptable.

In contrast, when we interpret (9b) the variants are the same but the restriction is stronger. We restrict to those variants where y is 0. The only such variant is $(x = n, y = 0)$, in which case xy is 0, so we predict that y is 0 is sufficient for xy is 0 and hence that (9b) is acceptable (given that it also satisfies the other requirements of *because*).

This illustrates how the present approach captures failures of sufficiency. This is important since, as discussed in section 3.2.1, if we had instead defined the modal horizon as the most similar (or minimally different) worlds where the cause argument is true, then when that sentence is actually true we would expect the modal horizon to only contain the actual world. In section 4.4.1 we will see that a similar result holds in Kratzer's semantics of conditionals.

3.8 The boundless imagination

One of the most striking features of the imagination is its boundlessness. We can imagine all sorts of bizarre eventualities: teleportation, a rabbit appearing in a top hat, a switch spontaneously turning off. So far we have presented a general recipe to change a state to allow a sentence to become true. What we would like is a blanket guarantee that whenever a sentence is true at some world – any world, no matter how remote – then regardless of which world we happen to find ourselves in, we can imagine a world where that sentence is true. Our imagination should be free to roam the outermost reaches of logical space, if needed, without being tripped up by actuality. In formal terms, this boundlessness of the imagination is captured by the following principle.

- (54) **Boundless imagination principle.** For any set of worlds P , moment t , world w and sentence A , if A is true at some nomically possible world, then A is true at some world in $mh_{P,t}(w, A)$.

The boundless imagination principle ensures that whenever we interpret a conditional $A > C$ or sufficiency claim $A \gg C$ where A is nomically possible, $mh_{P,t}(w, A) \cap |A|$ will be nonempty, so the claim will not be trivial.

It turns out the boundless imagination principles are guaranteed to hold given the following principle, which we call *modularity*, together with some mild auxiliary assumptions. Let us say that a state s *determines* a sentence A just in case A is true in every nomically possible world containing s . Let us also say that s *properly determines* A just in case s determines A and there is a nomically possible world where A is true and a nomically possible world where A is false.

- (55) **Modularity.** For any state s and sentence A , if s properly determines A then s overlaps a state A is about.

Let me briefly comment on the intuitive meaning of modularity. I think of it as a locality axiom. It relates a global notion – proper determination – which is defined in terms of worlds, with a local notion: overlap. In other words, proper determination is a long-distance relation while overlap is a close relation. This is similar to how locality is understood in physics. For instance, Newton’s theory of gravity is non-local since gravity acts instantaneously at a distance, while Einstein’s theory respects locality since gravity acts through the local curvature of spacetime. Modularity is a locality axiom in the sense that it requires a certain kind of distant relation to imply a certain kind of close relation.

We adopt the word ‘modular’ from its use in engineering. A smartphone, for example, is called modular when one can change parts of it while leaving the rest intact. Modularity in this sense is stronger than simply having individually identifiable parts. For example, with Apples phones we can identify the battery, the motherboard, the screen, and so on, but this is not enough for Apple phones to count as modular, since their parts are only compatible with very specific parts from the same company. In the same way, for a state space to count as modular it is not enough that we can identify separate states. Modularity requires that the states interact with aboutness in particular way. To illustrate, let us imagine that each nomically possible world is an existing smartphone.⁴⁰ Let s to be an Apple charging port. Then given Apple’s proprietary design, s determines A , that the screen is also made by Apple. And since there are phones not made by Apple, s properly determines A . Then for the phone to count as modular according to our modularity axiom, when we imagine the phone with a different screen, we have to imagine the charging port changing too. Since we can intuitively imagine changing the screen but not the charging point, Apple phones do not count as modular on our definition – agreeing with what engineers have said about Apple phones.

The boundless imagination principle follows from modularity together with the following auxiliary assumptions.

(56) **Auxiliary assumptions.**

- a. Parthood principles: bounded completeness and no emergent parts.
- b. Negation invariance: every sentence is about the same states as its negation.
- c. The intervention time t is set to before A becomes true:
if A is true at w' then it is also true at $w_{\prec t} \frown w'_{\succeq t'}$ for any worlds w, w' , sentence A and A -variant t' of t .

These are all quite plausible. We already met the parthood principles in section

⁴⁰Of course, our true concept of nomic possibility is broader than this. A phone can be nomically possible without currently existing. We artificially restrict the notion of nomic possibility here to illustrate the failure of modularity. Actually, I have not found any natural cases where modularity fails. This raises the interesting question whether modularity is a conceptual truth.

3.3.3, where they were needed to ensure fusions are well-behaved. Negation invariance also holds under the analysis of aboutness in section 3.7.1 and chapter 4. The third condition says that when a sentence is true at a world, it remains true when we glue on the actual past up to intervention time. This is plausible given a ban on trivial intervention times, discussed in section 3.6.4.

3.8.1. PROPOSITION. *Every imaginative structure satisfying the conditions in (56) and modularity satisfies the boundless imagination principle.*

PROOF. We use the single-state definition of A -variants from section 3.6.1. Pick any state s , sentence A , intervention time t and set of nomic possibilities P . Suppose further that there is a nomically possible world where A is true. Either $t - A$ exists or it does not.

Suppose $t - A$ exists. Recall from section 3.6.1 that $t - A$ is the largest part of t that does not overlap the A -part of t . It follows from the definition that $t - A$ does not overlap any state A is about. For suppose it did: some state u is part of both $t - A$ and a state A is about. Then u is one of the A -parts of t , and is therefore part of $A\text{-part}(t)$. But then $t - A$ overlaps $A\text{-part}(t)$, contradicting the definition of $t - A$. Then $t - A$ does not overlap any state A is about. Since A and $\neg A$ are about the same states, $t - A$ does not overlap any state $\neg A$ is about. Then by modularity (55), $t - A$ does not properly determine $\neg A$: either (i) A is true at every nomically possible world, (ii) A is false at every nomically possible world, or (iii) s does not determine $\neg A$.

(i) Suppose A is true at every nomically possible world. Since the actual world is nomically possible, A is true at $w = w_{\prec t} \frown w_{\succeq t}$, and $w_{\prec t} \frown w_{\succeq t} \in mh_{P,t}(w, A)$ since t is an A -variant of itself. Case (ii) is ruled out by our assumption that A is true at some nomically possible world. (iii) So suppose s does not determine $\neg A$: there is a nomically possible world w' containing $t - A$ where A is true. Then $t - A$ is part of t' for some $t' \in w'$, so $w_{\prec t} \frown w'_{\succeq t'} \in mh_{P,t}(w, A)$. And since the intervention time is set to before A is settled, A is also true at $w_{\prec t} \frown w'_{\succeq t'}$.

Now suppose $t - A$ does not exist. Since parthood is bounded complete and has no emergent parts, by Fact 3.6.2, every moment is an A -variant of t . Then since there is a nomically possible world w' where A is true, for any moment $t' \in w'$, we have $w_{\prec t} \frown w'_{\succeq t'} \in mh_{P,t}(w, A)$ and A is true at $w_{\prec t} \frown w'_{\succeq t'}$, so there is a world in $mh_{P,t}(w, A)$ where A is true. \square

Modularity is a novel constraint, so let us spend some time becoming familiar with it. Consider the state space in Figure 3.30a, with three colour states (completely red, completely green, completely blue; depicted by paintbrushes), two shape states (square, circle), and each combination of colour and shape. This state space represents our own understanding of colour and shape: colour and shape can be freely combined, but shapes cannot be combined with each other (square circles do not exist), and nothing is both completely one colour and completely another colour.

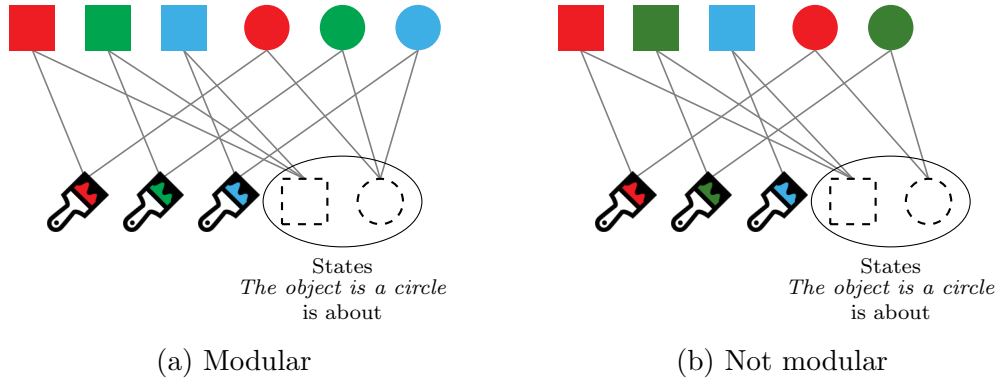


Figure 3.30

Suppose we see a blue square, and are asked to imagine it as a circle. We can imagine the shape changing while keeping the colour fixed. Let us assume that *The object is a circle* is about the shape states: the circle and square. In the state space of Figure 3.30a everything proceeds smoothly: the blue state does not overlap the square state, so when we imagine the blue square as a circle, we keep its colour.

Now let us see what happens when modularity fails. Imagine a state space without the blue circle state, given in Figure 3.30b. This state space declares blue circles to be impossible while leaving all other states intact. Assuming that *The object is a circle* is still about the set of shape states, modularity fails here because the blue colour state determines that the object is square, but it does not overlap any state *The object is a circle* is about.

Let us see how our proposal for imagining changes works in the state space of Figure 3.30b. Assuming that *The object is a circle* is about the circle and square, the blue paintbrush does not overlap any state this sentence is about. So when we imagine the blue square as a circle, the blue state remains. This state determines that the object is a square, so we would predict that it is impossible to imagine the blue square as a circle. Here the failure of modularity leads to a breakdown of imagination. Restoring modularity – either by adding the blue circle back in, or saying *The object is a circle* is about the blue state – restores the boundlessness of the imagination; imagining the blue square as a circle becomes possible again.

For a final illustration of the need for modularity, recall our discussion of the relation between colour and opacity in section 3.5. Suppose we are looking at a red stained glass window and consider:

(57) If the window were perfectly transparent, it would still be red.

This is intuitively unacceptable. If the object were perfectly transparent, it would not even have a colour. So when we imagine the window transparent, we do not fix its colour. To predict this on the present approach we need the state of the

window's colour to overlap a state *The window is perfectly transparent* is about.

In section 3.5 we proposed the following state space representing colour and opacity. Let $(O = n, C = m)$ be the state of the window's colour and opacity.

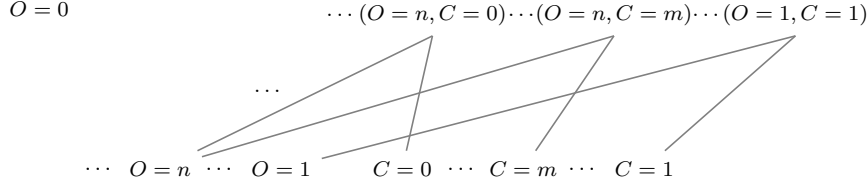


Figure 3.31

In this state space, the state $C = m$ properly determines $A = \textit{The window is not perfectly transparent}$. Modularity then requires that $C = m$ overlap a state A is about. Since A and $\neg A$ are about the same states, *The window is perfectly transparent*, this state space and modularity together correctly predict that when we imagine the window perfectly transparent, we do not fix its colour.

Compare this to what happens in the state space of Figure 3.32, generated by random variables, which we argued does not represent our intuitive concept of opacity and colour. Now $C = m$ does not properly determine A , since $C = m$

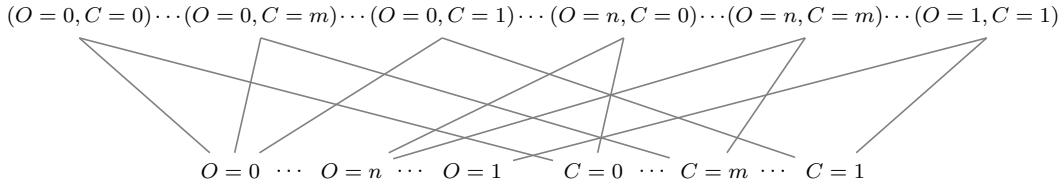


Figure 3.32

is part of the state $(O = 0, C = m)$ where A is false. In this state space, then, modularity allows that when we imagine the window perfectly transparent, we predict that we fix its colour. This undesirably allows (57) to come out true, once again illustrating the pitfalls of generating the set of states from random variables.

3.9 Conclusion

This essay has been an attempt to understand the systematic nature of the imagination. We observed that our cognition furnishes us with the ability to vary a part of the world while leaving all else intact, and saw that this ‘all else’ can be captured by parthood: it consists of the parts of the world that do not overlap the part we wish to vary (section 3.3). The framework is sufficiently general to represent dependence between properties, unlike approaches using random variables

such as structural causal models, and does not need to assume logical atomism (section 3.4).

We then extended this proposal to consider the hypothetical scenarios raised by a conditional antecedent or causal claim (section 3.6), which provided us with an analysis of sufficiency.

Chapter 4

Imagining logically complex sentences true

It is plausible to assume that the foreground of a logically complex sentence is not determined at random, but in a systematic way.¹ In section 3.7.1 we analysed the foreground via two components: nomic relevance and exact relevance. We defined nomic relevance but not exact relevance. A natural idea, one we develop in this chapter, is that exact relevance is determined in a *compositional* way.

We begin with a language generated by a set of atomic sentences, negation, conjunction and disjunction. In this section we will consider two ways to extend the exact relevance to logically complex sentences, which I call the *truthmaker semantics* view and the *subject matter* view.

- (1) A state s is exactly relevant to a sentence A just in case ...
Truthmaker semantics view: s exactly verifies or exactly falsifies A .
Subject matter view: s is in the subject matter of A .

Following Fine (2017b), we assume we have a function that assigns to each atomic sentence the set of its exact verifiers, and a function that assigns to each atomic sentence the set of its exact falsifiers. On both views of the foreground, a state s is exactly relevant to an atomic sentence just in case s exactly verifies or exactly falsifies s . They differ in how they define exact relevance for logically complex sentences.

The truthmaker semantics view. We adopt the following exact verification and falsification clauses from Fine (2017b:561–563).

- $(\neg)^+$ s exactly verifies $\neg A$ iff s exactly falsifies A ;
 $(\neg)^-$ s exactly falsifies $\neg A$ iff s exactly verifies A ;

¹A brief discussion of some of the issues from this chapter (in particular sections 4.1.3 and 4.2) has been previously published in McHugh (2022).

- $(\wedge)^+$ s exactly verifies $A \wedge B$ iff s is the fusion of an exact verifier of A and an exact verifier of B ;
- $(\wedge)^-$ s exactly falsifies $A \wedge B$ iff s exactly falsifies A or exactly falsifies B ;
- $(\vee)^+$ s exactly verifies $A \vee B$ iff s exactly verifies A or exactly verifies B ;
- $(\vee)^-$ s exactly falsifies $A \vee B$ iff s is the fusion of an exact falsifier of A and an exact falsifier of B .

The subject matter view. For our purposes, the key difference between the subject matter of a sentence and its exact verifiers and falsifiers is that we take subject matter to be invariant under the logical connectives: A and $\neg A$ have the same subject matter, as do $A \wedge B$ and $A \vee B$ (Yablo 2014:42, Fine 2016:11, Berto 2018:1878). We will adopt what Hawke (2018) calls an ‘atom-based’ approach to subject matter, whereby the subject matter of a sentence is the subject matter of the atomic sentences it contains.

- (2) a. A state is in the subject matter of an atomic sentence p iff it exactly verifies or exactly falsifies p .
- b. A state is in the subject matter of a complex sentence A iff it is in the subject matter of an atomic sentence in A .

Since the subject matter of a sentence is determined by the subject matter of its atoms, subject matters are automatically invariant under the logical connectives.







4.1 Conditional inference patterns






In this section we give a taste of the interaction between the present framework and the clauses we have just introduced, by showing that the present approach, on both the truthmaker semantics and subject matter views, invalidates some commonly-discussed inference patterns of conditional logic.

4.1.1 Invalidating antecedent strengthening

Antecedent strengthening is inference from $A > C$ to $A^+ > C$, where A^+ entails A . An instance of this rule is the inference from $A > C$ to $(A \wedge B) > C$. The failure of antecedent strengthening was a key success of ordering approaches to conditionals (e.g. Lewis 1973b). The present approach also invalidates antecedent strengthening. To illustrate, recall the light switch set up of Figure 3.1 where the light is on just in case both switches are up. Currently switch A is down and B is up. Consider (3).

- (3) a. If switch A were up, the light would be on.
- b. If switch A were up and B were down, the light would be on.

Intuitively (3a) is true and (3b) is false. Here is how we derive the truth of (3a) and the falsity of (3b) on the present approach. Let A be the sentence *switch A is up* and B be the sentence *switch B is down*. The actual state w is , and the only exact falsifier of A is , so the remainder is . The A -variants of w are the worlds containing this remainder:  and . Of these, the only A -variant of w where A is true is . In this world the light turns on, so we predict (3a) to be true, as desired.

Turning to (3b), when $A \wedge B$ is false, a state is in the foreground of $A \wedge B$ just in case it is in the foreground of A or it is in the foreground of B .  exactly falsifies A and  exactly falsifies B , so the $A \wedge B$ -variants of w are those worlds containing everything that does not overlap  or . (This is the same whether we adopt the truthmaker semantics or subject matter view.) In our toy state space of Figure 3.18, every part of the actual state  overlaps one of these states, so every world whatsoever counts as an $(A \wedge B)$ -variant of w . When we restrict to those where $A \wedge B$ is true, we find that in this world the light does not turn on, so (3b) is predicted to be false.

What is going on, informally speaking, is that $A \wedge B$ is ‘about’ more of the world than A . $A \wedge B$ has a larger foreground than A , so when we interpret $(A \wedge B) > C$, we allow more of the world to vary than when we interpret $A > C$. Conversely, A has a larger background than $A \wedge B$, so more of the world is kept fixed when we vary $A \wedge B$ than A . Allowing more of the world to vary gives rise to more variants: $A \wedge B$ broadens our imaginative horizons more than A .

4.1.2 Rational monotonicity

Rational monotonicity, also known as strengthening with a possibility, comes in two forms: one involving conditionals with existential modals such as *could*, which we will symbolise by $\Diamond \rightarrow$; the other in terms of *would*-conditionals, symbolised by $>$.

$$\frac{A > C \quad A \Diamond \rightarrow B}{(A \wedge B) > C} \qquad \frac{A > C \quad \neg(A > \neg B)}{(A \wedge B) > C}$$

Let us first consider the rule with $\Diamond \rightarrow$. Boylan and Schultheis (2017, 2021) offer the following counterexample to rational monotonicity.²

Alice, Billy, and Carol are playing a simple game of dice. Anyone who gets an odd number wins \$10; anyone who gets even loses \$10. The die rolls are, of course, independent. What Alice rolls has no effect on

²This example was inspired by a counterexample Stalnaker (1994) gave to rational monotonicity in belief revision.

what Billy rolls and vice versa. Likewise for Alice and Carol as well as for Billy and Carol.

Each player throws their dice. Alice gets odd; Billy gets even; Carol gets odd.

Consider the following sentences in this scenario.

- (4) a. If Alice and Billy had thrown the same type of number, then at least one person would still have won \$10.
 b. If Alice and Billy had thrown the same type of number, then Alice, Billy, and Carol could have all thrown the same type of number.
 c. If Alice, Billy, and Carol had all thrown the same type of number, then at least one person would still have won \$10.

Intuitively, (4a) and (4b) are acceptable while (4c) is not. This shows that rational monotonicity is intuitively invalid.

We can predict these judgements on the present proposal. First we need a semantics of $\Diamond \rightarrow$. The most natural extension of our proposal to conditionals with existential modals is given in (5), which we adopt here.

- (5) **Semantics of $\Diamond \rightarrow$.** Where P be the set of nomically possible worlds and t the intervention time, $A \Diamond \rightarrow C$ is true at a world w just in case there is a world in $mh_{P,t}(w, A) \cap |A|$ where C is true.

On this entry $\Diamond \rightarrow$ is dual not to *will/would*-conditionals, but to sufficiency. It is the selection function that breaks the duality between $>$ and $\Diamond \rightarrow$, while at the same time making $>$ self-dual, as shown by the dualities and implications in Figure 4.1 (assuming $mh_{P,t}(w, A) \cap |A|$ is not empty).

UNIVERSAL		SELECTIONAL		EXISTENTIAL
$A \gg C$	\Rightarrow	$A > C$	\Rightarrow	$A \Diamond \rightarrow C$
\Updownarrow		\Updownarrow		\Updownarrow
$\neg(A \Diamond \rightarrow \neg C)$	\Rightarrow	$\neg(A > \neg C)$	\Rightarrow	$\neg(A \gg \neg C)$

Figure 4.1

Duality of $>$ and $\Diamond \rightarrow$ fails whenever there is a world in $mh_{P,t}(w, A) \cap |A|$ where C is true, but the selection function happens to choose one where C is false. We do not want to add a selection function to the semantics of *can/could*. To see this, consider the following pair (based on an example from Higginbotham 2003).

- (6) a. Every coin would have landed heads if you had flipped it.

- b. No coin would have landed tails if you had flipped it.

These seem to say the same thing: every coin is rigged to land heads. Given the equivalence of *heads* and *not tails*, the selection function ensures the equivalence of $\forall x(\text{flip } x > \text{heads } x)$ and $\neg\exists x(\text{flip } x > \text{tails } x)$, as desired. Now replace *would* with *could*:

- (7) a. Every coin could have landed heads if you had flipped it.
b. No coin could have landed tails if you had flipped it.

The equivalence disappears. (7a) says that every coin is not rigged to land tails, while (7b) says that every coin is rigged to not land tails. If every coin is fair, (7a) is true but (7b) false.

Now that we have a plausible semantics of *can/could*-conditionals, we turn to Boylan and Schultheis's scenario. Let us analyse *Alice and Billy threw the same type of number* as $(\text{Alice odd} \wedge \text{Billy odd}) \vee (\text{Alice even} \wedge \text{Billy even})$, and similarly for *Alice, Billy and Carol threw the same type of number* (in section 4.2.3 we will give a semantics that predicts this automatically, without stipulation). For the state space, we suppose we have a state for each outcome of each roll and the possible fusions thereof. For example, there is a state of Alice rolling a five, which we assume exactly verifies that Alice rolled odd and exact falsifies that she rolled even. For concreteness, we will use the subject matter view of the foreground (the truthmaker semantics view makes the same predictions here).

On the subject matter view, the foreground of $A = (\text{Alice odd} \wedge \text{Billy odd}) \vee (\text{Alice even} \wedge \text{Billy even})$ is the set of exact verifiers/falsifiers of its atomics. The state of Carol's throw does not overlap any state in this foreground, and so is part of all worlds in the modal horizon. Focusing only on whether the rolls land odd or even, let w_{EEO} , say, be the world where Alice and Bob roll even and Carol rolls odd, and so on for the other rolls. Then

$$\begin{aligned} mh_{P,t}(w, A) \cap |A| &= \{w_{\text{OOO}}, w_{\text{OEO}}, w_{\text{EOO}}, w_{\text{EEO}}\} \cap \{w_{\text{OOO}}, w_{\text{OOE}}, w_{\text{EEO}}, w_{\text{EEE}}\} \\ &= \{w_{\text{OOO}}, w_{\text{EEO}}\}. \end{aligned}$$

In both of these worlds Carol still rolls odd, so someone still wins \$10 and we predict (4a) to be true. And since in at least one of these worlds everyone throws odd, we also predict (4b) to be true.

In contrast, the foreground of *Alice, Billy and Carol all threw the same type of number* contains the state of Carol's throw, so allow it to vary. The modal horizon expands to include the world where all three roll odd:

$$\begin{aligned} mh_{P,t}(w, A \wedge B) \cap |A \wedge B| &= \{w_{\text{OOO}}, w_{\text{OOE}}, w_{\text{OEO}}, w_{\text{OEE}}, w_{\text{EOO}}, w_{\text{EOE}}, w_{\text{EEO}}, w_{\text{EEE}}\} \cap \{w_{\text{OOO}}, w_{\text{EEE}}\} \\ &= \{w_{\text{OOO}}, w_{\text{EEE}}\} \end{aligned}$$

so we predict (4c) to be not necessarily true and therefore unassertable (since its truth depends on the selection function; see section 3.2.3). This shows how the present approach invalidates rational monotonicity with $\Diamond \rightarrow$, as desired.

Let us now turn to rational monotonicity with a negated conditional. The set $mh_{P,t}(w, A) \cap |A| = \{w_{OOO}, w_{EEO}\}$ contains a world where Alice, Billy and Carol all throw the same type of number and a world where they do not, so the truth of $\neg(A > \neg C)$ depends on the selection function. Validity is truth preservation on all interpretations, so to invalidate rational monotonicity it is enough to find *some* interpretation where the premises are true but the conclusion false; that is, some selection function where $A > C$ and $\neg(A > \neg B)$ hold but $(A \wedge B) > C$ does not. Take, for instance:

$$s(w, \underbrace{\{w_{OOO}, w_{EEO}\}}_{mh_{P,t}(w, A) \cap |A|}) = w_{EEO} \qquad s(w, \underbrace{\{w_{OOO}, w_{EEE}\}}_{mh_{P,t}(w, A \wedge B) \cap |A \wedge B|}) = w_{EEE}$$

Rational monotonicity fails here since B and C are true at w_{EEO} but C is false at w_{EEE} . So both forms of rational monotonicity are invalid on the present proposal.

On the ordering approach to conditionals, rational monotonicity corresponds to *almost connectedness*: for all worlds w, x, y, z , if $x <_w z$ then $x <_w y$ or $y <_w z$. Boylan and Schultheis (2021) show how to construct orders that are not almost connected in a principled way, following Kratzer (1981a). So rational monotonicity is easy to invalidate on ordering approaches to conditionals – just drop almost connectedness. In this respect, then, the present proposal is on a par with ordering approaches: both easily invalidate rational monotonicity.

In the next section we consider a rule closer to the heart of ordering approaches to conditionals: cautious monotonicity.

4.1.3 Cautious monotonicity

Cautious monotonicity is the inference

$$\frac{A > B \quad A > C}{(A \wedge B) > C} \text{ Cautious monotonicity}$$

To test its status, consider the set up in Figure 4.2. There are two switches, A and B, connected to a light. Part of the image is shaded. Each switch has three possible positions: up, in the middle, or down. As the wiring indicates, the light is on just in case A is in the middle and B is either up or in the middle. Currently, A is in the middle and B is down, so the light is off. Consider (8) in this context.

- (8) a. If switch B were in the shaded region, the light would be on.
- b. If switch B were in the shaded region, both switches would be in the shaded region.

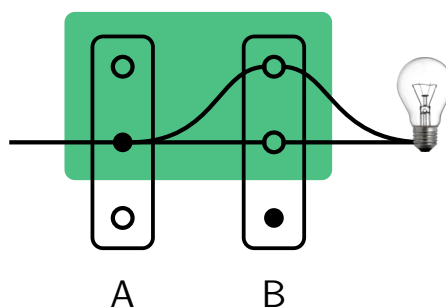


Figure 4.2: The light is on just in case switch A is in the middle and switch B is either in the middle or up. Currently A is in the middle and B is down, so the light is off.

- c. If both switches were in the shaded region, the light would be on.

Intuitively, (8a) and (8b) are clearly acceptable, but (8c) is dubious. The interpretation of (8c) is subtle and would certainly benefit from empirical testing. Nonetheless, one does not want to say that (8c) follows from (8a) and (8b) as a matter of logic. But if cautious monotonicity is valid, (8a) and (8b) together entail (9).

- (9) If both switches were in the shaded region and switch B were in the shaded region, the light would be on.

This is a roundabout way of saying (8c), assuming that *(both A and B) and B* and *(both A and B)* are equivalent in conditional antecedents (conditional antecedents are strange, but surely not so strange that this equivalence fails).

If cautious monotonicity were valid, there would be consequences beyond the acceptability judgements in (8). For instance, valid inference preserves probability and hence certainty: if *A* entails *B*, the probability of *B* cannot be less than the probability of *A*; in particular, if *A* has probability 1, *B* must too. If cautious monotonicity were intuitively valid, we would expect the probability of (8c) to be at least as high as the probability of the conjunction of (8a) and (8b). It seems, however, that (8a) and (8b) could both have probability 1 while (8c) has probability less than 1. Cautious monotonicity is not cautious enough. We need a framework with the flexibility to allow the rule to fail.

The present framework provides that flexibility. This is because in this framework conditional antecedents and consequents play fundamentally different roles. The job of a conditional antecedent is to point to a part of the world to vary, which is sensitive to non-truth-conditional factors such as what the sentence is about. In contrast, the job of a conditional consequent is simply to provide truth conditions.

To see how we can predict a reading where (8c) is unacceptable, let us assume the obvious state space for this scenario, with a state for each position of each

switch, one for each setting of the light, and the possible fusions thereof. One way to predict a reading where (8c) is unacceptable in the present framework is to assume that the state of both switches is exactly relevant to *both switches are in the shaded region*. Given that this state, $A \text{ MIDDLE} \sqcup B \text{ DOWN}$, is nomically relevant to the sentence, our definition of the foreground in (47) implies that it is in the foreground of $A \wedge B$. Since $A \text{ MIDDLE}$ and $B \text{ DOWN}$ both overlap $A \text{ MIDDLE} \sqcup B \text{ DOWN}$, they are allowed to vary. When we restrict to worlds containing the remainder where both switches are up, we find among them a world where A is up and the light off, therefore predicting (8c) to be unassertable.

It turns out that both the truthmaker semantics and subject matter views of the foreground predict (8c) to be true. To derive this prediction, we assume that *both switches are in the shaded region* has the logical form $A \wedge B$, that the state of switch A being in the middle exactly verifies *Switch A is in the shaded region* and that the state of switch B being down exactly falsifies *Switch B is in the shaded region*. The subject matter view in addition requires assuming the nomic relevance condition from section 3.7.1.³ If it turns out there is robust evidence for a false reading of (8c), then, one may wish to modify the exact relevance clauses of the truthmaker semantics and subject matter views of the foreground (or alternatively, in the case of the subject matter view, abandon the nomic relevance condition).

4.1.4 Reciprocity

There is a close connection between cautious monotonicity and the following rule, which Nute (1980b) calls ‘CSO’ and Egré and Rott (2021) call ‘reciprocity’.

$$\frac{A > B \quad B > A}{(A > C) \leftrightarrow (B > C)} \text{ Reciprocity}$$

³Let us derive these predictions here. On the truthmaker semantics view, we have to find the exact verifiers or falsifiers of *Both switches are in the shaded region*. Since this sentence is actually false, we find its exact falsifiers. According to truthmaker semantics’ clauses, a state exactly falsifies $A \wedge B$ just in case it exactly falsifies A or exactly falsifies B (or exactly falsifies $A \wedge B$; this extra condition does not affect predictions in this case). The state of switch A being in the middle does not exactly falsify that switch A is in the shaded region (nor, of course, does it exactly falsify that switch B is in the shaded region). So on the truthmaker semantics view the state of switch A being in the middle is not exactly relevant to $A \wedge B$. Since we assumed exact relevance is necessary to be in the foreground, this state is not in the foreground of $A \wedge B$. As the state of switch B being down exactly falsifies B , the foreground of $A \wedge B$ contains only the state of switch B being down: we do not vary the position of switch A and so predict (8c) to be true.

On the subject matter view, a state is exactly relevant to $A \wedge B$ just in case it is exactly relevant to A or exactly relevant to B . The state of switch A being in the middle exactly verifies that switch A is in the shaded region, so it is exactly relevant to A and hence exactly relevant to $A \wedge B$. However, this state is not nomically relevant to $A \wedge B$, so if we assume the nomic relevance condition in (47) it is not in the foreground of $A \wedge B$ and we still predict (8c) to be true.

Our scenario also serves as a counterexample to reciprocity.

- (10) a. If B were in the shaded region, both switches would be in the shaded region. $B > (A \wedge B)$
 b. If both switches were in the shaded region, B would be in the shaded region. $(A \wedge B) > B$
 c. If B were in the shaded region, the light would be on. $B > C$
 d. If both switches were in the shaded region, the light would be on. $(A \wedge B) > C$

(10a)–(10c) are all acceptable, but as discussed, (10d) is dubious.⁴

It is not surprising that a counterexample to cautious monotonicity is also a counterexample to reciprocity. For given basic principles of conditional logic – namely, identity ($A > A$), conjunction ($(A > B) \wedge (A > C) \rightarrow A > (B \wedge C)$) and the fact that $>$ is weaker than entailment – reciprocity implies cautious monotonicity.⁵

$$\frac{\frac{\overline{A > A} \text{ Identity} \quad A > B}{A > (A \wedge B)} \text{ Conjunction} \quad \frac{\overline{(A \wedge B) > B} \quad A > C}{(A \wedge B) > C} \text{ Reciprocity}$$

So where cautious monotonicity fails, reciprocity does too.

4.1.5 Attempting to invalidate cautious monotonicity on the ordering approach

The scenario we have just discussed illustrates a point often made (e.g. by Veltman 1976, 2005, Pollock 1976, Tichý 1976, Fine 1975a, 2012b), that the semantics of counterfactuals is not based on our intuitive concept of similarity. Consider

⁴Gabbay (1972:101) also rejects reciprocity, offering the following purported counterexample.

A = I am elected president of the U.S.
 B = I am recalling the U.S. troops from Asia.
 C = I am nicely dressed.

It may be true that if I were elected president I would have recalled the U.S. troops from Asia [$A > B$], also if I were to recall the U.S. troops from Asia I would be elected president [$B > A$], and certainly if I am elected president I am nicely dressed [$A > C$]; but that does not imply that if I were to recall the troops from Asia I would be nicely dressed [$B > C$].

I do not find this to be a convincing counterexample. Unfortunately Gabbay does not expand on why he finds $B > C$ unacceptable in this context.

⁵Indeed, if we also assume cautious transitivity, cautious monotonicity implies reciprocity:

$$\frac{\frac{A > B \quad A > C}{(A \wedge B) > C} \text{ Cautious monotonicity} \quad B > A}{B > C} \text{ Cautious transitivity}$$

the configurations in Figure 4.3, with the actual configuration on the left, and answer the questions in (11).

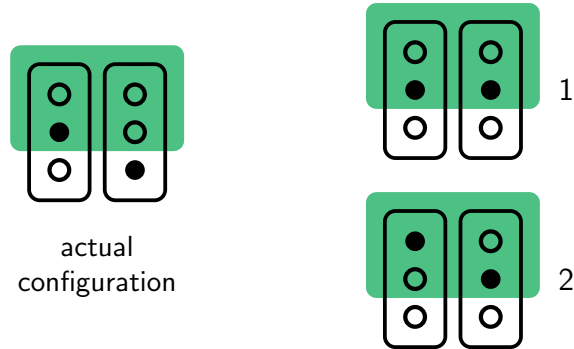


Figure 4.3

- (11) a. Is configuration 1 more similar than configuration 2 to the actual configuration?
 b. Is configuration 2 more similar than configuration 1 to the actual configuration?

Intuitively, the answers are ‘Yes’ and ‘No’, respectively. Or, putting things in terms of minimal difference (which is the intuitive notion underlying the accounts of Stalnaker 1968 and Pollock 1976), answer the questions in (12).

- (12) a. Is configuration 1 one where both switches are in the shaded region, and which otherwise differs minimally from the actual configuration?
 b. Is configuration 2 one where both switches are in the shaded region, and which otherwise differs minimally from the actual configuration?

Here the natural answers are ‘Yes’ and ‘No’, respectively. After all, configuration 1 differs from the actual configuration only in the position of switch B, while configuration 2 differs from the actual configuration in the position of both switches. If we are looking for the most similar worlds to the actual world where both switches are in the shaded region, configuration 2 involves, in the words of Lewis (1979), a ‘gratuitous difference’. So if the semantics of counterfactuals were based on our intuitive concept of similarity or minimal difference, we would expect (8c) to be as acceptable as the claim that configuration 1 is more similar to the actual configuration than configuration 2. But this is not what we observe.

Since (8c) is not clearly true, one might try to rescue a semantics of counterfactuals based on ordering over worlds by offering a different intuitive gloss of the ordering, one not based on our intuitive concept of similarity or minimal difference. Stalnaker, for example, writes that “the relevant conception of minimal difference needs to be spelled out with care” (Stalnaker 1984:129). The problem,

however, runs deeper than that. Cautious monotonicity is part of the *logic* of all accounts of the semantics of counterfactuals based on an ordering over worlds, including those of Stalnaker (1968) and Lewis (1973b). On Lewis's semantics, for example, we have:

- (13) a. \leq is a function from worlds to binary relations over worlds.
 b. $A > C$ is true at w with respect to \leq iff for every A -world v there is an A -world $u \leq_w v$ such that C is true at every A -world $u' \leq_w u$.

On this semantics, cautious monotonicity follows from reflexivity and transitivity of \leq_w .⁶ Reflexivity and transitivity are the absolute minimum constraints to impose on the order; without them the ordering approach would be deeply unworkable. Moreover, if we construct the order from an ordering source as Kratzer (1981b:47) proposes, given in (14),

- (14) a. An ordering source g is a function from worlds to sets of propositions.
 b. $u \leq_w v$ just in case for all $p \in g(w)$, if p is true at v , p is true at u .

then reflexivity and transitivity of the order follow, respectively, from the reflexivity and transitivity of implication – bedrock principles of logic.

The point is that cautious monotonicity is hardwired into ordering approaches to conditionals.⁷ One might wish to respond to (8) by saying that there is something funny going on with salience or attention. But the semantics in (13) precludes such responses. No matter how much care we take to spell out the order, cautious monotonicity will come out valid. (8) shows that the logic of ordering approaches to counterfactuals is too rigid.

4.2 Simplification

Simplification (of disjunctive antecedents) is the rule

$$\frac{(A \vee B) > C}{(A > C) \wedge (B > C)} \text{ Simplification}$$

⁶Let us quickly prove this. Suppose $A > B$ and $A > C$ are true at w and pick any $(A \wedge B)$ -world v . Since $A > C$ is true at w , there is an A -world $u \leq_w v$ such that every A -world $u' \leq_w u$ is a C -world. And since $A > B$ is true at w and u is an A -world, there is an A -world $x \leq_w u$ such that $A \rightarrow B$ holds at every $x' \leq_w x$. By reflexivity, A and B are true at x . And for any $x' \leq_w x$, since $x \leq_w u$, by transitivity of \leq_w , $x' \leq_w u$, so if x' is an A -world, it is a C -world. A fortiori, if x' is an $(A \wedge B)$ -world, it is a C -world.

⁷That being said, Delgrande (1987) proposes a logic NP for nonmonotonic inference that invalidates cautious monotonicity by adopting a semantics other than (13). NP is intended to capture statements about prototypical properties (e.g. statements of the form *normally As are Bs*) rather than conditionals.

This is a rich literature on simplification.⁸ The present account does not validate this rule. One mundane reason for this is the selection function: there is no constraint forcing the selected $(A \vee B)$ -world to be the same as the selected A -world. So C can be true at the selected $(A \vee B)$ -world but false at the selected A -world, which allows $(A \vee B) > C$ to be true while $A > C$ is false. However, our task is to model our linguistic behaviour, where the operative notion is arguably not truth but assertability. The more interesting question, therefore, is not whether simplification preserves truth but whether it preserves assertability. In section 3.2.3 we proposed that a conditional is assertable only if it is true on all selection functions. A conditional $A > C$ is true on all selection functions just in case A is sufficient for C , so asking whether simplification preserves assertability amounts to asking whether simplification with \gg in place of $>$ preserves truth.

$$\frac{(A \vee B) \gg C}{(A \gg C) \wedge (B \gg C)} \text{ Simplification for sufficiency}$$

The present approach does not validate this rule either. This is a welcome result, since there are intuitive counterexamples to simplification, such as (15) from McKay and Inwagen (1977).

- (15) If Spain had fought with the Allies or the Axis, they would have fought with the Axis.

From (15) we certainly do not infer (16).

- (16) If Spain had fought with the Allies, they would have fought with the Axis.

(15) is assertable while (16) clearly is not, which shows that simplification does not preserve assertability.

Now what, intuitively, does (15) say? This is reasonably clear: it says that World War II Spain had strong Axis sympathies. They were greatly disposed to join the Axis over the Allies. By including this preference in our model, we can account for the unacceptability of (15), as we see now.

Let us take the intervention time t to be a moment when Spain was considering entering the war. Simplifying greatly, we represent Spain at t using two states: Axis preference and neutrality, which we take to be disjoint, depicted in Figure 4.4. The larger state space in which t finds itself, again greatly simplified, is given in Figure 4.5.

Turning to nomic possibility, there are three relevant rules: if Spain is neutral, it does not join the war; if Spain is not neutral and prefers the Axis, it joins the

⁸Among the authors who claim simplification is valid are Nute (1975), Ellis, Jackson, and Pargetter (1977), Warmbröd (1981), Fine (2012b), Starr (2014), and Willer (2018). Among those who claim it is invalid are Nute (1980b, 1984), Bennett (2003), van Rooij (2006), Santorio (2018), and Lassiter (2018).

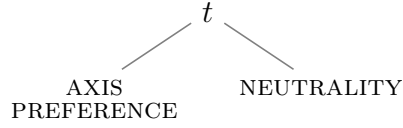


Figure 4.4

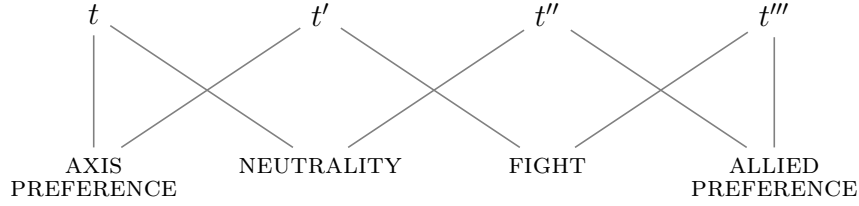


Figure 4.5

Axis; and if it is not neutral and prefers the Allies, it joins the Allies. Given this, we place the following constraints on the set of nomically possible worlds.

- (17) **Nomic possibility.** For any nomically possible world w ,
- a. if NEUTRALITY is part of a moment in w , Spain does not join the war in w ;
 - b. if AXIS PREFERENCE and FIGHT are both part of a moment in w , Spain joins the Axis in w ;
 - c. if ALLIED PREFERENCE and FIGHT are both part of a moment in w , Spain joins the Allies in w .

We assume the state of Spain's neutrality and the state of its Axis preference are exactly relevant to $A = \textit{Spain fights with the Allies}$. Since both states are also nomically relevant to A , they are in the foreground of A .

Turning to $A \vee B$, since both NEUTRALITY and AXIS PREFERENCE are exactly relevant to A , they are also exactly relevant to $A \vee B$, by the exact relevance clause for disjunction on both the truthmaker and subject matter views of the foreground. Also, NEUTRALITY is nomically relevant to $A \vee B$, since it determines that the sentence is false: if Spain is neutral they fight with neither the Axis nor the Allies. However, AXIS PREFERENCE is not nomically relevant to $A \vee B$. This is because Spain can have an Axis preference while staying neutral, in which case $A \vee B$ is false, and they can have an Axis preference and fight with the Axis, in which case $A \vee B$ is true; also, they can have no Axis preference and stay neutral, in which case $A \vee B$ is false, and they can have no Axis preference and fight with the Allies, in which case $A \vee B$ is true.

Finally, we assume that the larger states, t, t', t'' and t''' are not in the foreground of $A \vee B$ because they are not exactly relevant to $A \vee B$ (notice that they each have a proper part that is exactly relevant to $A \vee B$, which we assume is

	nominally relevant to $A \vee B$	exactly relevant to $A \vee B$
NEUTRALITY	✓	✓
AXIS PREFERENCE	✗	✓

enough to violate exact relevance in this case). So our foregrounds are:

$$\begin{aligned} \text{foreground}(A \vee B) &= \{\text{NEUTRALITY}\} \\ \text{foreground}(A) &= \{\text{NEUTRALITY}, \text{AXIS PREFERENCE}\} \end{aligned}$$

AXIS PREFERENCE does not overlap any state in the foreground of $A \vee B$, while NEUTRALITY does. So when we interpret $(A \vee B) > C$ we allow Spain's neutrality to vary while fixing its Axis preference. Given the state space of Figure 4.5, the $A \vee B$ -variants of t are t itself and t' . But with (16), Spain's Axis preference and neutrality are both allowed to vary when we interpret (16). Every moment in our model is therefore an A -variant of t .

$$\begin{aligned} A \vee B\text{-variants}(t) &= \{t, t'\} \\ A\text{-variants}(t) &= \{t, t', t'', t'''\} \end{aligned}$$

Given the nomic possibilities, the possible continuations of t are worlds where Spain stays out of the war, while the possible continuations of t' are worlds where Spain joins the Axis. When we restrict the modal horizon to worlds where $A \vee B$ is true, we find that in all such worlds Spain joins the Axis. So we correctly predict (15) to be true. While when we restrict to worlds containing an A -variant of t where Spain fights with the Allies, we find that, of course, they fight with the Allies, not the Axis, so we correctly predict (16) to be false.

This example illustrates an important feature of how we defined the foreground: the separate requirement of both exact relevance and nomic relevance. One might think that whenever a state is exactly relevant to a sentence, it is nominally relevant to the sentence too. This is not the case. To see this, note that exact relevance is preserved under disjunction – whenever a state is exactly relevant to A it is exactly relevant to $A \vee B$ – while nomic relevance is not. For example, Spain's Axis preference is nominally relevant to *Spain fights with the Allies* but not *Spain fights with the Allies or the Axis*. It follows that exact relevance does not imply nomic relevance. Spain's Axis preference is exactly relevant to *Spain fights with the Axis or Allies* but is not nominally relevant to this sentence. Now, the nomic relevance constraint is quite plausible. It seems required by the intuition that when we are asked to imagine a sentence true or false, we only vary states that are responsible for the sentence having the truth value it has. For example, intuitively Spain's Axis sympathy was not responsible for them staying out of the war, so this idea implies that when we imagine them joining the war, we fix the fact that they had Axis sympathy. Since a state can be exactly relevant to a sentence while being nominally irrelevant to it, requiring exact relevance alone

in the definition of the foreground it not enough. We need the nomic relevance constraint too.

As stated, many semantic theories of conditionals validate simplification. A prominent reply to McKay and Inwagen's example on behalf of simplification's validity is that when (15) we judge true, we assume that Spain joining the Axis is in some sense not genuinely possible (Warmbröd 1981, Starr 2014, Willer 2018, Fine 2012b).⁹ However, Lassiter (2018) observes that simplification can fail even when we regard both disjuncts as possible, as shown in (18) (for further examples see Lassiter 2018).

- (18) If Spain had fought with the Axis or the Allies it's likely, but not certain, that they would have fought with the Axis.

Lassiter puts *likely* in the consequent, but the judgement pattern still stands when we put it in different positions or outside the sentence altogether.

- (19) It's likely, but not certain, that if Spain had fought with the Axis or the Allies they would have fought with the Axis.
- (20) a. Alice: If Spain had fought with the Axis or the Allies, they would have fought with the Axis.
b. Bob: I think that's likely, but not certain.

One who utters (18), (19) or (20a) clearly takes Spain fighting with the Allies to be a counterfactual possibility. Now, given standard assumptions about the meaning of *likely*, whenever *A* entails *B* and *A* is likely, *B* is likely too (see Yalcin 2010:921). Then if simplification were valid, (18), (19) and Bob's response in (20) would imply that (16), repeated below, is likely true – an undesirable result.

- (16) If Spain had fought with the Allies, they would have fought with the Axis.

McKay and Inwagen's original example in (15) intuitively communicates that Spain had a strong preference to join the Axis. We modelled this by assuming

⁹Starr (2014:1049) argues that the infelicity in the following sentence confirms that, for (15) to be accepted, there cannot be any accessible worlds in which Spain fights with the Allies.

- (i) Spain didn't fight on either the Allied or Axis side, but she could have ended up with the Allies. #Nevertheless, if she had fought for the Axis or the Allies, she would have fought for the Axis.

An alternative response is that the utterance of *Spain could have ended up with the Allies* adds to the modal horizon worlds where Spain fights with the Allies, and that these are carried over when we interpret the conditional in (i). There is independent evidence that previously mentioned possibilities can enter the domain of later modals (see e.g. von Fintel 2001b), which we can account for using von Fintel's modal horizon, or modal subordination as Starr (2014) proposes. See note 9 for a formalisation of the present account using the modal horizon.

that in *every* nomically possible world containing AXIS PREFERENCE and FIGHT, Spain joins the Allies. (18) instead communicates a softer preference to join the Axis. We can model this by adjusting the nomic possibilities accordingly. Suppose that among the nomically possible worlds containing AXIS PREFERENCE and FIGHT, there are some where Spain joins the Axis and some where Spain joins the Allies, and that restricted to the nomically possible worlds containing AXIS PREFERENCE and FIGHT, the probability of Spain joining the Axis is greater than 50% but less than 100%.

Given these nomic possibilities, AXIS PREFERENCE is no longer in the foreground of $A = \textit{Spain joins the Allies}$. Given how we defined nomic relevance, Spain merely having a soft preference for the Axis is not nomically relevant to them joining the Allies. Now the foregrounds of $A \vee B$ and A are the same; namely, $\{\text{NEUTRALITY}\}$, so the $A \vee B$ -variants and A -variants of t are also the same: $\{t, t'\}$. When we restrict to the worlds containing an $A \vee B$ -variant of t where $A \vee B$ is true, we find the probability of joining the Axis is greater than 50% but less than 100%, so in this model (18) is true (assuming a standard meaning for *likely* and *certain*, together with a way to relate selection functions to probabilities, such as the proposal by Cariani and Santorio 2018:§8). But when we restrict to the worlds containing an A -variant of t where A is true, we find that in all of them Spain fights with the Allies, so in this model (16) is false, and hence we do not predict it to be likely true.

The present approach therefore accounts for intuitive failures of simplification, both in the original counterexamples and those where both of the antecedent's disjuncts are considered possible.

4.2.1 Against Universal Realisability of the Antecedent

In the same way we accounted for failures of simplification, we can handle the following examples, which Embry (2014) gave against the semantics of counterfactuals in Fine (2012b). Fine proposes a principle he calls *Universal Realisability of the Antecedent*: if a counterfactual is true then it is true for any way in which its antecedent is true. Embry (2014) offers the following counterexamples.

- (21) a. If it had snowed yesterday, I would have gone skiing.
b. If it had snowed 100 feet yesterday, I would have gone skiing.
- (22) a. If Sue were to take some of these pills, she would get better.
b. If Sue were to take 25 of these pills, she would get better.
- (23) a. If Sue were not in the driver's seat, she would have survived.
b. If Sue were in between the two colliding cars, she would have survived.

The challenge is to account for why, when we interpret the (a)-sentences, we typically do not consider the scenarios raised by the (b)-sentences. Fine (2012b)

proposes that $A > C$ is true at world w just in case for every exact verifier t of A and possible outcome u of t at w , u inexactly verifies C . It is natural to suppose that there is a possible state where it snowed 100 feet that exactly verifies *It snowed*, and that this state has a possible outcome where the speaker does not ski. Given this, Fine's semantics predicts (21a) to be false. Fine (2012a:232) may respond that when take (21a) to be true, we do not regard it snowing 100 feet yesterday as 'genuinely possible'. But then to make predictions about how we interpret conditionals we would need an account of how we decide which scenarios count as 'genuine possibilities'.

In contrast, on the present approach we can account for these cases in the same way we accounted for failures of simplification. Take (21). We may assume that part of yesterday's weather conditions determined that it will not snow 100 feet; say, the clouds in the vicinity were not carrying enough water for that to happen. Another part of the weather conditions, disjoint from this state, was responsible for it not snowing at all yesterday – say, the fact the temperature was above freezing. There is a part of the scenario determining that it does not snow 100 feet disjoint from every state in the foreground of *It snowed yesterday*. Thus the conditions determining that it does not snow 100 feet are held fixed when we interpret (21a). Under these quite reasonable assumptions, we predict that when we interpret (21) we do not consider scenarios where it snows 100 feet.

To put this account to the test, imagine that instead yesterday's weather conditions in fact favoured it snowing 100 feet yesterday. Actually it didn't snow, but meteorologists discovered that the clouds contained so much water, if the weather conditions for snow had been present, there would have been 100 feet of snow. Under that assumption our judgement of (21a) changes, exactly as we expect on this account.

Similarly, we can account for (22) by assuming that Sue, say, has a desire to follow her doctor's instructions, and that this determines that she does not take 25 pills. The state of this desire does not overlap the state of her not taking any pills. (We know they do not overlap since there are possible scenarios containing one but no part of the other.) So when we remove the fact that she did not take any pills, the desire remains, and this desire ensures – encoded via nomic possibility – that if Sue takes some pills, she only takes the prescribed amount.

It is not surprising that we account for failures of simplification in the same way we account for Embry's cases; for we can draw a direct parallel between the two. In event semantics it is typically assumed that free variables denoting events are interpreted as existentially quantified, via an operation called existential closure (Davidson 1967a, who traces the idea to Reichenbach 1947). For example, *It snowed yesterday* can be interpreted as $\exists e(e \text{ is a snowing event} \wedge \text{runtime}(e) \subseteq \text{yesterday})$. Given that existential quantification is generalised disjunction, we may write a generalised form of simplification as the inference from $(\exists x Ax) > C$ to $\forall x(Ax > C)$. Failures of simplification show that this rule fails where x ranges over sentences and Ax states that x is in the relevant domain and

true. Embry's cases show that this rule fails where x ranges over events and Ax states that x is in the relevant domain and actual. Given this parallel, Embry's examples helpfully show that failures of simplification are a special case of the failure of a broader logical pattern, one not unique to disjunction.

4.2.2 Cases supporting simplification

So far we have only considered counterexamples to simplification. Let us now turn to cases supporting its validity. We first consider the example that kicked off discussions about simplification, due to Nute (1975).

- (24) If we had had good weather this summer or the sun had grown cold, we would have had a bumper crop.

(24) is intuitively unacceptable. Let us assume that this sentence has the logical form $(A \vee B) > C$ (in the next section we will explore alternative readings). Nute designed (24) as a counterexample to similarity analyses of counterfactuals. For it is natural to assume that having good weather requires less of a departure from the actual world than the sun growing cold (more precisely, for every world where the sun grows cold, there is a world strictly more similar to the actual world where we have good weather this summer). Given this, similarity-based approaches to conditionals (such as Stalnaker 1968, Lewis 1973b) undesirably predict (24) to be equivalent to (25).

- (25) If we had had good weather this summer, we would have had a bumper crop.

This problem does not arise for our proposal because it is not based on the concept of similarity. Let us show how our approach can account for (24)'s unacceptability.

We assume that the world at intervention time t contains the following two states: WEATHER NOT GOOD and SUN STAY WARM, representing, respectively, the fact that the weather was not good and that the sun stays warm. We assume that WEATHER NOT GOOD is exactly relevant to $A = \textit{we had good weather this summer}$ and that SUN STAY WARM is exactly relevant to $B = \textit{the sun grew cold}$. Then both states are exactly relevant to $A \vee B$.

Regarding nomic possibility, we assume that for every nomically possible world w , the weather is not good in w just in case w contains WEATHER NOT GOOD, and the sun stays warm in w just in case w contains SUN STAY WARM. Then by our definition of nomic relevance, both states are nomically relevant to $A \vee B = \textit{We had good weather this summer or the sun grew cold}$, since the absence of either state determines the truth of $A \vee B$. So both states are in the foreground of $A \vee B$ and therefore are allowed to vary when we interpret (24). There is some moment not containing SUN STAY WARM (there is some nomically possible world where the sun grows cold). This moment is an $(A \vee B)$ -variant of t , and every

world containing it is one where the sun grows cold, in which case we do not have a bumper crop. Thus $mh_{P,t}(w, A \vee B) \cap |A \vee B|$ contains a world where the consequent is false, so we predict (24) to be unassertable.

The key difference between McKay and Inwagen's example and Nute's is that, while WWII Spain was disposed to join the Axis over the Allies, there is no state like GOOD WEATHER PREFERENCE OVER SUN GROWING COLD such that for every nomically possible world containing this state, if the weather is good or the sun grows cold in this world, the sun stays warm. While it is easy to imagine WWII Spain harbouring Axis sympathy, it is bizarre to imagine a part of the world manifesting a preference between good weather and the sun growing cold.

Given the plausible counterexamples to simplification, semantic theories of conditionals have given many proposals to account for cases where the rule is felt to preserve truth. For example, Alonso-Ovalle (2006, 2009), Ciardelli (2016), and Ciardelli, Zhang, and Champollion (2018) assume that disjunction introduces alternatives and that conditionals involve universal quantification over alternatives. Such a proposal is compatible with the present account of the modal horizon, and certainly has appeal; for example, it accounts for the behaviour of disjunction in conditionals and unconditionals in a uniform way (see Rawlins 2013). However, the proposal raises the question of where the universal quantification comes from. It is not expected from the semantics of *would*, which as we saw in 3.2.2, does not contribute universal quantification.¹⁰ And as we will see below, putting universal quantification over alternatives into the literal meaning of conditionals makes the wrong predictions for conditionals in downward entailing environments. For these reasons, in the following section we aim to find the source of the universal quantification over alternatives within the semantic entry for conditionals we assumed in (18).

Or-to-and. Here is a surprising feature of the present approach: in many cases the following rule will come out to be truth-preserving.

$$\frac{(A \vee B) > C}{(A \wedge B) > C} \text{ Or-to-and}$$

There are compelling counterexamples to this rule, as in the following scenario from Schulz (2007:105) (also discussed by Ciardelli, Zhang, and Champollion 2018). Imagine two switches connected to a light. Each switch can be up or down. The light is on just in case both switches are in the same position. Currently both switches are up. Consider:

¹⁰Of course, one may easily rewrite the semantics of $A > C$ to include universal quantification; say, as $\forall w' \in \{s(mh_{P,t}(w, A) \cap |A|)\}, w' \in |C|$, and then argue that the universal quantification over alternatives is somehow derived from this universal quantifier. But this feels like a trick, one that ignores the possibility of seeing simplification as an instance of general principles governing natural language interpretation.

- (26) a. If switch A or B were down, the light would be off.
 b. If switch A and B were down, the light would be off.

The present account predicts $(A \text{ down} \vee B \text{ down}) > \text{Light off}$ to be unassertable in this scenario. Simply put, this is because on the present account when we interpret this antecedent we ‘remove’ the states of both switches, which leaves room for both switches to be down. This possibility survives the restriction to worlds where A or B is down (with *or* read inclusively).¹¹ But in this world the light is on.

We propose that the apparent invalidity of or-to-and and the apparent validity of simplification have the same source. More concretely, we propose that the reading observed for (26a) is a free choice inference: disjunction takes wide-scope above the whole conditional, resulting in $(\text{if } A, C) \vee (\text{if } B, C)$ which is strengthened to a conjunctive interpretation $(\text{if } A, C) \wedge (\text{if } B, C)$ as a free choice inference (Figure 4.6). On this account, (26a) is acceptable because it is

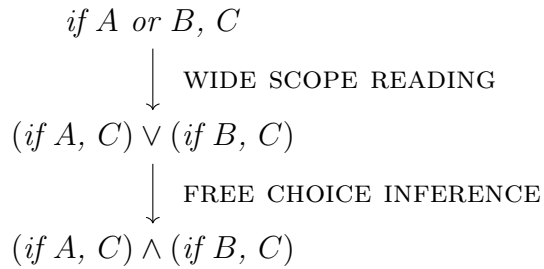


Figure 4.6

interpreted as $(\text{if } A, C) \wedge (\text{if } B, C)$.

Our key assumption to get this proposal off the ground is the following.

- (27) Wide scope constraint. $[(A > C) \vee (B > C)]$ is preferred to $[(A \vee B) > C]$.

An important question facing this proposal, one we will not answer here, is why disjunctive antecedent conditionals prefer to be interpreted with wide scope disjunction. There is precedent for items with a disjunctive/existential meaning preferring wide scope and being strengthened by a free choice inference. For example, Dayal (2004) and Chierchia (2013:333) propose that free choice *any* is subject to a wide scope constraint and receives its universal meaning by a free choice inference – exactly parallel to the proposal here. A fruitful question to explore is whether the wide scope constraint for universal free choice items can also account for the wide scope reading of disjunction in conditionals of the form *if A or B, C*.

¹¹For evidence that *or* is by default read inclusively in disjunctive conditional antecedents, see Ciardelli, Zhang, and Champollion 2018:§2.6.2

As we will see, each step in this account is independently attested: *if A or B, C* can be read as $(\text{if } A, C) \vee (\text{if } B, C)$, and conditionals overtly of this form are strengthened to $(\text{if } A, C) \wedge (\text{if } B, C)$ by default.

This proposal gives a straightforward account of why simplification sometimes fails.¹² In general, when an expression has a scope ambiguity and its default resolution is unavailable, we try another resolution. In cases where simplification fails the wide scope free choice reading $(\text{if } A, C) \wedge (\text{if } B, C)$ is false. For example, if Spain had fought with the Allies they would not have fought with the Axis; if the sun had grown cold we wouldn't have had a bumper crop. This leaves the narrow scope reading $\text{if } (A \vee B), C$.¹³ In some cases this reading gives a perfectly apt utterance; for example, in section 4.2 we saw that McKay and Inwagen's example, on a narrow scope reading, successfully communicated that WWII Spain had Axis sympathies. In other cases both scope readings are out. As we saw in section 4.2, Nute's example (24) is also false on the narrow scope reading. The sentence has nowhere to turn, so we correctly predict its unacceptability.

We argue for a free choice approach to simplification on the basis of the many parallels between the two inferences.

Four analogies between free choice and simplification.

Analogy 1. Both free choice sentences and disjunctive conditional antecedents have a wide-scope reading without free choice/simplification.

- (28) You may have cake or ice cream. I can't remember which. (Fusco 2019, Pinton 2021)
- a. \nrightarrow You may have cake.
 - b. \nrightarrow You may have ice cream.
- (29) You would have gotten an extension if you had talked to the rector or the vice-rector. I can't remember which one is responsible for extensions.
- a. \nrightarrow You would have gotten an extension if you had talked to the rector.
 - b. \nrightarrow You would have gotten an extension if you had talked to the vice-rector.

¹²Other approaches that claim conditionals with disjunctive antecedents are ambiguous include van Rooij (2006), Alonso-Ovalle (2009), Starr (2014), Santorio (2018), and Khoo (2021b), though none account for simplification using the scope ambiguity that we use, between disjunction scoping inside the conditional antecedent or over the whole conditional.

¹³I assume that the wide scope disjunction reading of *if A or B, C* as $(\text{if } A, C) \vee (\text{if } B, C)$ without the free choice inference is only available when the speaker avows ignorance about which conditional is true. So it is not an available reading in the counterexamples to simplification we have considered here.

Analogy 2. There are instances of both wide-scope free choice (Zimmermann 2000) and wide-scope simplification.

- (30) You can have cake or you can have ice cream.
 - a. \leadsto You can have cake.
 - b. \leadsto You can have ice cream.
- (31) Alice might be in Paris or she might be in Berlin.
 - a. \leadsto Alice might be in Paris.
 - b. \leadsto Alice might be in Berlin.
- (32) If you cut the grass I'll give you 5 euro or if you clean the windows I'll give you 5 euro.
 - a. \leadsto If you cut the grass I'll give you 5 euro.
 - b. \leadsto If you clean the windows I'll give you 5 euro.

Wide scope simplification also appears in the Book of Leviticus:

- (33) And if a soul sin ... if he do not utter it, then he shall bear his iniquity.
 Or if a soul touch any unclean thing ... he also shall be unclean, and guilty.
 Or if he touch the uncleanness of man ... when he knoweth of it, then he shall be guilty. (Leviticus 5:1–3, [King James Version](#), 1611).

This is most naturally read as a conjunction of conditionals. Given that the book has been translated into many languages, this passage invites cross-linguistic comparison. A disjunction word links the clauses of Leviticus 5 in, for example, [Mandarin Chinese](#) (*huò*), the original [Hebrew](#) (*o*), [Hungarian](#) (*vagy*), [Icelandic](#) (*eða*), [Māori](#) (*rānei*), [Urdu](#) (*yâ*), [Somali](#) (*ama*), [Welsh](#) (*neu*) and [Yoruba](#) (*tàbí*). This suggests that wide-scope free choice from conditionals is a cross-linguistically robust phenomenon.

Analogy 3. Free choice inferences and simplification both disappear in downward entailing environments (without special ‘denial’ intonation).¹⁴

¹⁴One might wonder whether *doubt* could scope below the conditional antecedent at LF. Iatridou has found cases where the antecedent takes surface scope above an attitude verb:

- (i) If it rains Mary believes/said/heard/assumed that Bill will come. (Iatridou 1991:26)

One might try to devise a mechanism where in (35) and (36) the free choice is computed with the attitude verb in the consequent, proposing that (35) and (36) have following logical forms, respectively.

- (ii) a. If Alice or Bob had taught him, John doubts he would have passed.
 b. If Alice or Bob had taught her, every student doubts she would have passed.

I see two challenges facing this idea. The first is that (35) implies both (35a) and (35b) (and

(34) I doubt Alice can have ice cream or cake.

- a. \leadsto I doubt Alice can have ice cream.
- b. \leadsto I doubt Alice can have cake.

Alonso-Ovalle (2009, citing Angelika Kratzer, p.c.) and Santorio (2018) observe that, in downward entailing environments, conditionals with disjunctive antecedents also entail both of their simplifications.

(35) John doubts he would have passed if Alice or Bob had taught him.

- a. \leadsto John doubts he would have passed if Alice had taught him.
- b. \leadsto John doubts he would have passed if Bob had taught him.

(36) Every student doubts she would have passed if Alice or Bob had taught her.

- a. \leadsto Every student doubts she would have passed if Alice had taught her.
- b. \leadsto Every student doubts she would have passed if Bob had taught her.

(37) No one will enjoy the party if they get stuck talking to Larry or Sue.

- a. \leadsto No one will enjoy the party if they get stuck talking to Larry.
- b. \leadsto No one will enjoy the party if they get stuck talking to Sue.

Given that free choice inferences disappear in downward entailing environments, these readings follow immediately on the present account, since $\neg((\text{if } A, C) \vee (\text{if } B, C))$ is equivalent to $\neg(\text{if } A, C) \wedge \neg(\text{if } B, C)$.

Analogy 4. Free choice inferences and simplification both allow for a ‘denial reading’.

(36) implies both (36a) and (36b)) even when the attitude verb is interpreted below conditional antecedent; indeed, this is the most natural reading, where the doxastic states in question are not conditional on who teaches. The second problem is that the QP *every student* in (iib) does not c-command the pronoun *her* (see Iatridou 1991:30–31), so the sentence is not even grammatical.

A further potential piece of evidence against this strategy comes from (iii).

- (iii) A: If Alice or Bob had taught John, he would have passed.
- B: I doubt it.

B’s response intuitively implies that Alice and Bob are both bad teachers. However, I do not find this datum so convincing. One might reply, say, that in (iii) B’s beliefs are modally subordinate (in the sense of Roberts 1989) to the antecedent *if Alice or Bob had taught John*, as they are in (iv).

- (iv) A: If a wolf came in to the house, it would eat the apples.
- B: I doubt it would.

- (38) Alice cannot have cake OR ice cream. Ice cream is off limits.

Here is an example from the TV show *The Office* (season 2, episode 2).

- (39) Michael: The problem is that I am the boss, and apparently I can't say anything.
Jan: Well, that's true... in a way. You can't say ANYthing.

Here Jan is denying the free choice reading of *anything*.¹⁵

A hallmark of denial readings is that they require overt negation. Compare:

- (40) a. Alice is not allowed to have cake OR ice cream. Ice cream is off limits.
b. ??Alice is forbidden from having cake OR ice cream. Ice cream is off limits.

We see the exact same pattern with denials of the simplification inference, requiring the same intonation contour:

- (41) Samee does not think that he would have passed if Alice OR Bob had taught him. He thinks Bob is a great teacher.

and overt negation. Compare (41) with (42).

- (42) ??Samee doubts that he would have passed if Alice OR Bob had taught him. He thinks Bob is a great teacher.

¹⁵Alonso-Ovalle (2009) offers the following example to support the claim that simplification results from universal quantification over alternatives.

- (i) It is plain false that Hitler would have been pleased if Spain had joined Germany or the U.S. (Kratzer, p.c., a variation on an example in Nute (1980a:157))
(ii) ... There is enough evidence showing that he might have objected to Spain joining the U.S. If she had joined Germany, he would have been pleased, of course. (Kratzer, p.c.)

Alonso-Ovalle (2009:220) later writes that he will remain agnostic as to whether $\neg(\text{if } A \text{ or } B, C)$ implies $\neg(\text{if } A, C)$ and $\neg(\text{if } B, C)$ or $\neg(\text{if } A, C) \text{ or } \neg(\text{if } B, C)$.

I take it that (i) involves a denial reading, where the characteristic emphasis on *or* is obscured by the emphatic *it is plain FALSE that....* Compare the sentences in (iii).

- (iii) a. It is plain false that Alice would have enjoyed the party if her best friend or ex had come.
b. It is plain false that Alice would have enjoyed the party if her BEST FRIEND or ex had come.

(iiia) suggests that Alice does not enjoy spending time with her best friend, while (iiib) only suggests that Alice does not enjoy spending time with her ex.

Comparison with alternative accounts of simplification.

Santorio (2018) Santorio (2018) proposes that simplification is due to a covert distributivity operator, and that when the distributivity operator is absent, simplification fails. A problem for Santorio’s approach is that distributivity is optional, and is often absent when it leads to an incoherent meaning.

- (43) The students ate a pizza. ... They were still hungry afterwards, since they each had only one slice!

The students ate a pizza is most naturally interpreted with a covert distributivity operator, expressing that the students ate a pizza each. The continuation in (43) forces a non-distributive interpretation, showing that the distributivity operator is optional. Now, if (44) similarly contained an optional distributivity operator, would expect to be able to remove it to save the sentence from falsity.

- (44) If we had had good weather this summer or the sun had grown cold, we would have had a bumper crop. (Alonso-Ovalle 2006, a variation on an example by Nute 1975)

If simplification were due to a distributivity operator, as Santorio (2018) proposes, we would have to explain why it seems obligatory for conditionals but optional elsewhere.

Moreover, Khoo (2021a) noticed that simplification is obligatory with *if or if*-conditionals, suggesting they do not use an (optional) distributivity operator.

Khoo (2021) Khoo (2021a) considers the case of *double-if* conditionals, proposing that *if* can type-shift propositions to sets of propositions.

- (45) $\llbracket \text{If } A \text{ or if } B, C \rrbracket^{w,f} = 1 \text{ iff } \forall \mathbf{X} \in \{\mathbf{A}, \mathbf{B}\} : \forall w' \in f^{\mathbf{X}}(w) : \llbracket C \rrbracket^{w',f^{\mathbf{X}}} = 1.$

A problem with Khoo’s (2021) account is that it predicts an existential meaning for *if or if*-conditionals in downward entailing environments. We instead observe a universal meaning. To see this, let’s see what happens when we embed an *if or if*-conditional under *doubt*. *Doubt* is downward entailing, as shown by the fact that it licences NPIs (a fact that is expected assuming the equivalence of *doubt* and *think not*).

- (46) a. I doubt anyone will have any questions.
b. I doubt she ever tried pickles.
c. I doubt he lifted a finger to help.

Now consider (47), in a context where every student attended all the classes, was taught by Alice and passed.

- (47) Every student doubts they would have passed if they had missed a class or if Bob had taught them.

(47) has two readings, depending on the scope of *or*.

- (48) a. *every* > *or*. For every student x , x doubts that (x would have passed if x had missed a class or x would have passed if Bob had taught x).
 b. *or* > *every*. Every student doubts they would have passed if they had missed a class or every student doubts they would have passed if Bob had taught them.

We are interested in the first reading. It intuitively implies that for every student x , x doubts that they would have passed if they had missed a class **and** x doubts that they would have passed if Bob had taught them.

- (49) a. Every student x doubts $((\text{if } x \text{ miss, pass}) \vee (\text{if Bob teach } x, \text{pass}))$
 b. \Leftrightarrow Every student x thinks $\neg((\text{if } x \text{ miss, pass}) \vee (\text{if Bob teach } x, \text{pass}))$
 c. \Leftrightarrow Every student x thinks $(\neg(\text{if } x \text{ miss, pass}) \wedge \neg(\text{if Bob teach } x, \text{pass}))$

We observe the same strong reading with other downward entailing operators:

- (50) No student would have passed if they had missed a class or if Bob had taught them.
 a. \leadsto No student would have passed if they had missed a class.
 b. \leadsto No student would have passed if Bob had taught them.

Under downward entailing operators, then, Khoo (2021a) predicts an existential meaning for *if or if*-conditionals that at least one of the conditionals is false. We instead observe the universal meaning that both conditionals are false.

In contrast, on the present proposal $\neg(\text{if } A, C \text{ or if } B, C)$ gives rise to the observed universal meaning: $\neg(\text{if } A, C)$ and $\neg(\text{if } B, C)$, since free choice inferences disappear in downward entailing environments (denial readings aside).

4.2.3 First-order clauses

We may define exact verification and exact falsification for a first-order language. This will extend the reach of our analysis of sufficiency and conditionals to quantified sentences. For a conditional antecedent or *because*-clause containing a quantified sentence often raises hypothetical alternatives in a systematic way, with different people imagining the same scenarios. Since the hypothetical scenarios raised by a sentence are determined by its exact verifiers and falsifiers (and the parthood relation), to capture this fact on the present proposal we need the exact verifiers and falsifiers of quantified sentences to also be determined in a systematic way.

Our atomic sentences will be of the form $P(a_1, \dots, a_n)$ where P is an n -ary

predicate and a_1, \dots, a_n individual constants. In the first-order case our model is as follows.

- (51) **First-order model.** Our model is a tuple $(S, \leq, D, g, |\cdot|^+, |\cdot|^-)$ where
- a. (S, \leq) is a state space.
 - b. D is a domain of individuals.
 - c. g is a function assigning to each term an element of the domain.
 - d. $|\cdot|^+$ and $|\cdot|^-$ assign to each n -ary predicate P and sequence of n individuals (d_1, \dots, d_n) from the domain a set of states.

The clauses for the atomic sentences become:

- (52) a. $s \models P(a_1, \dots, a_n)$ iff $s \in |P, g(a_1), \dots, g(a_n)|^+$
 b. $s \models \neg P(a_1, \dots, a_n)$ iff $s \in |P, g(a_1), \dots, g(a_n)|^-$

and the clauses for the connectives are the same as in the propositional case.

There are many ways to extend the propositional clauses to the first order case (see Fine 2017b:566–569 for discussion). Here we will follow the tradition of Generalised Quantifier Theory (Barwise and Cooper 1981, Keenan and Stavi 1986) by representing quantifiers as having two arguments, a restriction and a scope (this approach is also discussed by Fine 2017b:568). Where $A(x)$ and $B(x)$ are sentences and x a variable, $\forall x(A(x) : B(x))$ and $\exists x(A(x) : B(x))$ are also sentences. To illustrate, the sentence *Every switch is up* can be formalised as $\forall x(\text{switch}(x) : \text{up}(x))$, where *switch* and *up* are unary predicates.

An immediate issue one runs into when formulating exact verification and falsification clauses for the quantifiers is the question of world-relativity. Since entities may have different properties in different worlds, it appears that a state may be present in two worlds and exactly verify a quantified sentence in one world but not in the other. For example, let w_2 be a world containing only two light switches, A and B, both of which are up, and w_3 a world containing switches A, B and a third switch C, which are also all up. Then the state of A and B being up is part of both w_2 and w_3 , but it is plausible to say that this state exactly verifies *every switch is up* in w_2 but not in w_3 .

World-relativity is important issue when it comes to giving the exact verification and falsification clauses for the quantifiers, but it is not directly relevant to our present goal here of describing what parts of the world we fix and what we allow to vary when interpreting a sentence.¹⁶ We will therefore take the simplest

¹⁶ Fine (2017b:568) proposes that the exact verifiers of a universal statement (and dually, the exact falsifiers of an existential statement) contain a ‘totality fact’. For instance, there is a state representing the fact that switches A and B are all the switches there are, which is part of w_2 but not w_3 . On such a proposal exact verification and falsification are not world-relative. For the proposal denies that the state of switch A and B being up exactly verifies *every switch is up* in w_2 ; rather, this sentence is exactly verified by the fusion of that state together with the totality fact representing that A and B are all the switches. In w_3 a different state – containing

possible view on the exact verification and falsification clauses for the quantifiers, assuming a fixed domain of entities and a fixed interpretation of the predicates across worlds. We say that a state s exactly verifies $\forall x(A(x) : B(x))$ just in case s is a fusion of exact verifiers $B(a_1), B(a_2), \dots$, where a_1, a_2, \dots are all and only the constants such that $A(a_1), A(a_2), \dots$ are true. Note that this clause blends two notions of verification: truth and exact verification. We have truth in the restrictor and exact verification in the scope. We take truth to be classical (e.g. $\neg A$ is true just in case A is not true, and so on). The exact verifiers of an existential statement are straightforward. A state s exactly verifies $\exists x(A(x) : B(x))$ just in case it exactly verifies $B(a)$ for some constant a such that $A(a)$ is true. Dually, we define the exact falsifiers of quantified sentences similarly, by swapping \forall and \exists , and swapping exact verification and exact falsification. Our semantic clauses are as follows.¹⁷

$(\forall)^+ s \Vdash \forall x(A(x) : B(x))$ iff there is a set of states T such that $s = \bigsqcup T$ and a function $f : C \rightarrow T$ where $C = \{c : c \text{ is a constant, } A(c) \text{ is true}\}$ and $f(c) \Vdash B(c)$ for all $c \in C$.

$(\forall)^- s \nVdash \forall x(A(x) : B(x))$ iff $s \nVdash B(a)$ for some constant a such that $A(a)$ is true.

$(\exists)^+ s \Vdash \exists x(A(x) : B(x))$ iff $s \Vdash B(a)$ for some constant a such that $A(a)$ is true.

$(\exists)^- s \nVdash \exists x(A(x) : B(x))$ iff there is a set of states T such that $s = \bigsqcup T$ and an assignment $f : C \rightarrow T$ where $C = \{c : c \text{ is a constant, } A(c) \text{ is true}\}$ and $f(c) \nVdash B(c)$ for all $c \in C$.

On the subject matter view of the foreground, we take the subject matter of a quantified sentence to percolate up from the subject matter of its instances for which the restriction holds. Then just as $A \vee B$ and $A \wedge B$ have the same subject matter of the states of switches A, B, C and a different totality fact – exactly verifies *every switch is up*.

¹⁷These clauses produce some familiar desirable results. Here we mention two. Firstly, universal and existential quantification come out as duals (let $A \equiv B$ denote that A and B have the same exact verifiers and falsifiers at every world).

$$\begin{aligned} \forall x(A(x) : B(x)) &\equiv \neg \exists x(A(x) : \neg B(x)) \\ \exists x(A(x) : B(x)) &\equiv \neg \forall x(A(x) : \neg B(x)) \end{aligned}$$

Secondly, universal and existential quantification are generalised conjunction and disjunction, respectively, in the following sense.

$$\begin{aligned} A(a) \wedge A(b) &\equiv \forall x(x = a \vee x = b : A(x)) \\ A(a) \vee A(b) &\equiv \exists x(x = a \vee x = b : A(x)) \end{aligned}$$

matter, $\exists x(A(x) : B(x))$ and $\forall x(A(x) : B(x))$ will have the same subject matter.

- (53) a. A state s is in the subject matter of $\exists x(A(x) : B(x))$ just in case it is in the subject matter of $B(a)$ for some constant a such that $A(a)$ is true.
 b. A state s is in the subject matter of $\forall x(A(x) : B(x))$ just in case it is in the subject matter of $\exists x(A(x) : B(x))$.

4.2.4 Putting formal conditions in the restrictor

In this section we discuss an advantage of distinguishing between a quantifier's restrictor and scope, and interpreting the restrictor with respect to ordinary truth at a world rather than exact verification and falsification. Doing so makes it much easier to determine the foreground of sentences whose meaning includes purely formal conditions, such as tense information. For example, in section 3.7.1 we discussed the sentence *This is moving*, proposing that once we attend to tense and aspect we can represent its meaning as

$$\exists e(\text{move}(e) \wedge \text{agent}(e) = x \wedge \text{runtime}(e) \subseteq t).$$

Now, our guiding intuition when analysing the foreground was that a sentence's foreground is the set of states that are 'directly responsible' for its truth. It is natural to wonder what part of the world is 'directly responsible' for the truth of a condition such as $\text{runtime}(e) \subseteq t$, stating that the runtime of event e is included in the reference time t , or $\text{agent}(e) = x$, stating that x is the agent of event e . However, given a distinction between a quantifier's restrictor and scope, we can instead represent the meaning of *This is moving* as

$$\exists e(\text{agent}(e) = x \wedge \text{runtime}(e) \subseteq t : \text{move}(e)).$$

To determine the foreground of this formula, we need to know the truth conditions of the restrictor and the foreground of the scope. We do not need to determine the foreground of the formal conditions $\text{agent}(e) = x$ and $\text{runtime}(e) \subseteq t$. The issue, of course, is not unique to *This is moving* but fully general. Virtually all sentences express formal conditions, such as information locating events in time. We can apply the present proposal without needing to answer tricky questions about the exact verifiers or falsifiers of such formal conditions – a welcome result.

Let us now apply the clauses to some example sentences, to test the predictions of our proposal.

4.2.5 Applying the first-order clauses

Consider a variant of the light switch example from Figure 3.1. This time there are three switches, A, B and C, and the light is on just in case all switches are up. Currently A is down and B and C are up, so the light is off (see Figure 4.7).

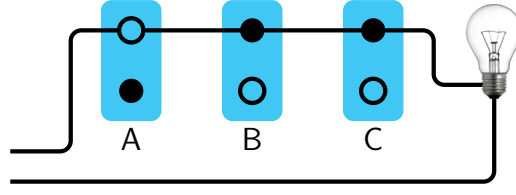


Figure 4.7: The light is on just in case all three switches are up.

Consider (54) in this context.

- (54) a. If all the switches were up, the light would be on.
 b. If the switches were in the same position, the light would be on.

Or suppose all the switches are up and consider (55).

- (55) a. The light is on because all the switches are up.
 b. The light is on because all the switches are in the same position.

Intuitively the (a)-sentences in (54) and (55) are true while the (b)-sentences are false.

Let us work out what we predict for this example on the truthmaker semantics view of the foreground (the subject matter view gives the same results here). We begin by computing the states in the switches scenario that exactly verify or falsify *all the switches are up*. Since this sentence is currently false, we find the exact falsifiers of this sentence.

$$\begin{aligned}
 & s \Vdash \forall x(\text{switch}(x) : \text{up}(x)) \\
 \text{iff } & s \Vdash \text{up}(d) \text{ for some constant } d \text{ such that } \text{switch}(d) \text{ is true at } w \quad (\forall)^- \\
 \text{iff } & s \Vdash \text{up}(a) \text{ or } s \Vdash \text{up}(b) \text{ or } s \Vdash \text{up}(c)
 \end{aligned}$$

Let us assume that for any switch x , a state exactly falsifies *switch x is up* just in case it exactly verifies *switch x is down*. The actual world contains the state of A being down, the state of B being up, and the state of C being up. Of these, only the state of A being down overlaps an exact falsifier of $A = \forall x(\text{switch}(x) : \text{up}(x))$. By our definition of A -variant in Definition 3.6.1, the A -variants of w are those containing the state of B being up and C being up; namely, the actual configuration, and the one where all switches are up. When we then restrict to those A -variants where A is true, we find that all switches are up. In that world, the light is on, so we predict (54a) to be true, as desired.

Turning to the (54b), while the sentence *The switches are in the same position* does not contain any quantifiers overtly, one may represent its meaning at a certain level of abstraction using quantifiers. For example, one may take it to be equivalent to *There is a position x such that for every switch y , y is in x* ; that is,

$$\exists x(\text{position}(x) : \forall y(\text{switch}(y) : y \text{ is in } x)).$$

We calculate the exact falsifiers of this sentence as follows.

- $s \Vdash \exists x(\text{position}(x) : \forall y(\text{switch}(y) : y \text{ is in } x))$
- iff there is a set of states T such that $s = \bigsqcup T$ and $f : C \rightarrow T$ where
 $C = \{c : c \text{ is a constant, } \text{position}(c) \text{ is true}\}$ and
 $f(c) \Vdash \forall y(\text{switch}(y) : y \text{ is in } c)$ for all $c \in C$.
- iff there is a set of states T such that $s = \bigsqcup T$ and an assignment $f : \{up, down\} \rightarrow T$
 where $f(up) \Vdash \forall y(\text{switch}(y) : y \text{ is up})$ and $f(down) \Vdash \forall y(\text{switch}(y) : y \text{ is down})$
- iff there is a set of states T such that $s = \bigsqcup T$ and $f : \{up, down\} \rightarrow T$
 where $f(up) \Vdash A \text{ down}$ or $g(up) \Vdash B \text{ down}$ or $f(up) \Vdash C \text{ down}$
 and $f(down) \Vdash A \text{ up}$ or $g(down) \Vdash B \text{ is up}$ or $f(down) \Vdash C \text{ is up}$.

Where A DOWN, B UP and C UP are the states exactly verifying *A is down*, *B is up* and *C is up*, respectively, we see that we have two choices for T at w : $\{A \text{ DOWN}, B \text{ UP}\}$ and $\{A \text{ DOWN}, C \text{ UP}\}$. Thus their fusions $A \text{ DOWN} \sqcup B \text{ UP}$ and $A \text{ DOWN} \sqcup C \text{ UP}$ exactly falsify that the switches are in the same position. When we remove the states overlapping these states, we find that we remove the position of each switch. This is illustrated in Figure 4.8, where a , b and c abbreviate A DOWN, B UP and C UP, respectively, and, for example, ab abbreviates the fusion of a and b . We let r represent the part of the world that does not overlap the switches.

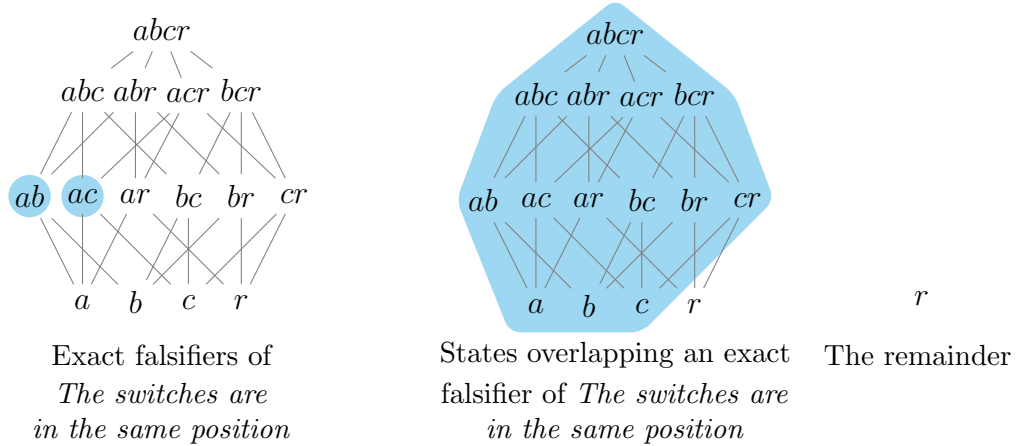


Figure 4.8

So where A is the sentence *The switches are in the same position*, there are A -variants where all switches are up and A -variants where all switches are down. In the latter world the light is off, so we predict that (54b) is not necessarily true,

since it is false for some selection functions, and therefore unassertable. This is the intuitively correct result.

Here is a second example, involving infinity.¹⁸ Imagine an infinite tape, where each cell of the tape can be in one of two states: with a hole or no hole. Actually, every cell has a hole (Figure 4.9).



Figure 4.9

Imagine a counting machine that runs along the tape, counting the number of holes. The counter has enough memory to count to one trillion. Suppose the machine starts counting. It counts to one trillion and then stops. It was not able to count all the holes. Consider (56) in this context.

- (56) The counter was unable to count all the holes because there are infinitely many of them.

(56) is an acceptable thing to say in this context, given that the counter can only count to one trillion.¹⁹

Recall the negative condition of *because* is an existential condition: it is satisfied if (but not only if) there is some possibility in the modal horizon where the cause occurs but the effect does not.²⁰ Some evidence for this comes from the acceptability of sentences such as (57) (McHugh 2020).

- (57) He has an American passport because he was born in Boston.²¹

Similarly, for the negative condition of *because* to be satisfied in (56), it is enough that the modal horizon contain *some* possibility where the tape has less than a trillion holes. Intuitively we consider such possibilities when we interpret (56).

This point is reinforced when we consider the conditional in (58).

- (58) If the tape had finitely many holes, it would still have more than a trillion holes.

¹⁸Similar issues involving infinity and verification have been discussed by Kratzer (1990, 2012:171), Armstrong (2004:21–22, citing Greg Restall) and Deigan (2020).

¹⁹Note that (56) may give rise to the inference that the counter was unable to count all the holes *only* because there are infinitely many of them, i.e. that for any finite number of holes, the counter would have been able to count them. On this reading (56) is unacceptable. In McHugh (2020) I argue that this is an implicature.

²⁰We write “but not only if” since this is not a necessary condition for the negative condition to be satisfied, as overdetermination cases make clear. But it is a sufficient condition. This follows from the factivity of production: since *C produce E* entails *E*, $\neg(C \gg E)$ entails $\neg(C \gg (C \text{ produce } E))$.

²¹Source: [Rupaulsdragrace.fandom.com](https://www.rupaulsdragrace.fandom.com).

(58) is unacceptable. On hearing it, one wants to respond: ‘not necessarily’. If the tape had finitely many holes, intuitively it could have any finite number of holes.

We can express *The tape has infinitely many holes* in a first-order language as *There is a collection x (plurality, set, etc.) such that x is infinite and for every y in x , y is a hole*:

$$\exists x(\text{collection}(x) \wedge \text{infinite}(x) : \forall y(y \in x : y \text{ is a hole})).$$

When we compute the exact verifiers of this sentence using the clauses proposed above, we find it is exactly verified by every fusion consisting of infinitely many holes.²² Our definition of A -variance asks us to consider the set of worlds containing every state that does not overlap an exact falsifier of the sentence we are asked to imagine true; namely, $A = \textit{The tape has finitely many holes}$. So the A -variants of the tape are those with any number of holes. When we then restrict to those A -variants where there are finitely many holes, we find tapes with any finite number of holes, matching intuition. In some of these worlds there are fewer than one trillion holes, which is enough to satisfy the negative condition of (56) and correctly predict its acceptability, and for (58) to come out as not necessarily true, and therefore unassertable (since its truth depends on the particular selection function).

4.3 Comparison with similarity approaches and premise semantics

Now that we understand how the present approach derives the correct predictions in these cases, let us take them as a point of comparison with alternative analyses of conditionals. We will consider two main approaches to conditionals, those based on similarity (Lewis 1973b) or minimal change (Stalnaker 1968, Pollock 1976), and Kratzer’s approach based on premise semantics (Kratzer 2012).

Consider the following three configurations of the three switches (Figure 4.10).

Now answer the following questions.

- (59) a. Is configuration 1 more similar than configuration 2 to the configuration on the left?

²²This agrees with a proposal by Deigan (2020:527), who suggests that a state exactly verifies *There are infinitely many F s* just in case it is the fusion of infinitely many states t, u, \dots where each state t, u, \dots exactly verifies $F(a)$ for some constant a . Deigan proposes this entry as a stipulation, while here it follows from general principles; in this case from $(\exists)^+$, and by assigning the sentence *There are infinitely many F s* the logical form $\exists x(\text{set}(x) \wedge \text{infinite}(x) : \forall y(y \in x : F(y)))$.

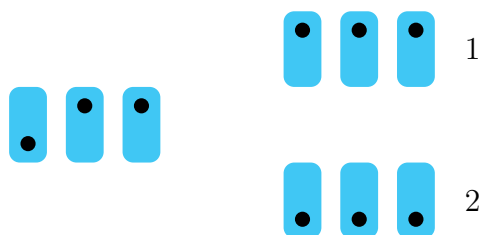


Figure 4.10

- b. Is configuration 2 more similar than configuration 1 to the configuration on the left?
- (60) a. Does configuration 2 differ more than configuration 1 from the configuration on the left?
- b. Does configuration 1 differ more than configuration 2 from the configuration on the left?

These questions do not have a single absolute answer. However, there is a prominent sense in which the answer to the (a)-questions is ‘yes’ and the answer to the (b)-questions is ‘no’. For configuration 1 differs from the configuration on the left only in the position of one switch, while configuration 2 differs in the position of two. In this sense 1 is more similar than 2 to the left configuration. This is certainly not the only way to answer these questions, but it is clear that according to our intuitive concept of similarity and difference, it is a plausible response – perhaps even the most plausible one.

Now recall (54b):

- (54b) If the switches were in the same position, the light would be on.

When we interpret (54b), intuitively we consider both configurations 1 and 2. Consider what a semantics of conditionals based on our intuitive concept of similarity or difference would say about this case. As we have seen, there is a plausible sense in which configuration 1 is more similar than 2 to the left configuration, and that 1 differs less than 2 from the left configuration. If our interpretation of conditionals were based on our intuitive concept of similarity or difference, we would expect this response to affect our interpretation of (54b), either by making (54b) come out true or by allowing us to acknowledge a prominent sense in which it is true. This is not what we observe. Rather, we reject (54b) as unassertable. This divergence suggests that our interpretation of conditionals is not based on our intuitive concept of similarity or difference.²³

The tape example raises similar problems for a semantics of conditionals based

²³A number of authors have given previous examples suggesting that the semantics of conditionals is not based on our intuitive concept of similarity or difference (see e.g. Fine 1975a, Tichý 1976, Slote 1978).

on similarity or difference. Imagine three tapes, A, B_1 and B_2 . Each tape has infinitely many cells, with each cell in one of two states: blank or with a hole punched out. As Figure 4.11 shows, every cell of tape A has a hole, tape B_1 has a hole only in the first cell, and tape B_2 has a hole only in the first two cells.

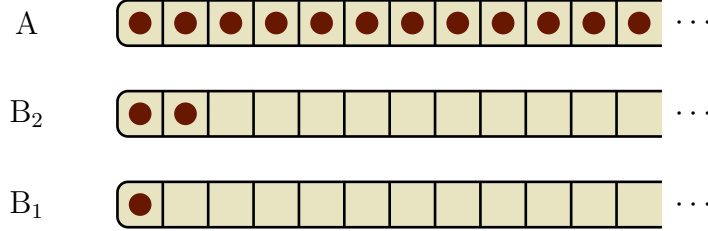


Figure 4.11

Let us evaluate how tapes B_1 and B_2 compare in terms of similarity to tape A. Consider:

- (61) a. Is B_1 more similar to A than B_2 is?
 b. Is B_2 more similar to A than B_1 is?

Intuitively, I would respond ‘No’ and ‘Yes’, respectively.

Of course, we can go on. Imagine a tape with holes in the first three cells (call it B_3). Intuitively B_3 more similar to A than B_2 . In general, for any natural number n , let B_n be the tape beginning with n many holes, and no holes thereafter, illustrated in Figure 4.12. According to our intuitive concept of similarity, for any natural number $n \geq 1$, B_{n+1} is more similar to A than B_n is.

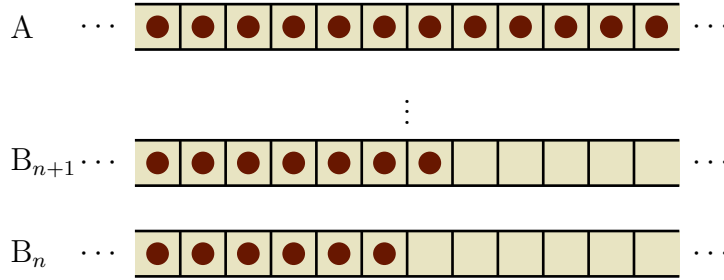


Figure 4.12

The set of tapes B-tapes, $\{B_n : 1 \leq n \in \mathbb{N}\}$ therefore does not contain a ‘most similar’ tape to A. Formally put, our intuitive concept of comparative similarity is not converse well-founded: the set of B-tapes does not have a maximal element with respect to similarity to tape A. Thus if we take Lewis’s (1973) semantics of *would* conditionals to be based on our intuitive concept of similarity, his semantics predicts (58), repeated below, to be acceptable.

- (58) If the tape had finitely many holes, it would still have more than a trillion holes.

Moreover, failures of the limit assumption lead to invalidating the following rule.

$$\frac{A > C_1 \quad A > C_2 \quad A > C_3 \quad \cdots}{A > (C_1 \wedge C_2 \wedge C_3 \wedge \cdots)} \text{Conjunction rule}$$

The conjunction rule is highly plausible. Now consider:

- (62) a. For every natural number n , if tape A had finitely many holes, it would have not have n holes.
 b. If tape A had finitely many holes, for every natural number n , the tape would not have n holes.
 c. If tape A had finitely many holes, it would not have finitely many holes.

Given that each tape B_{n+1} is strictly more similar to the actual tape A than B_n , Lewis's semantics for counterfactuals – based on our intuitive concept of similarity – would predict (62a) to be true. But then if the infinitary conjunction rule were valid, (62a) would imply (62b), which is equivalent to (62c), of the form $A > \neg A$. In other words, if we plug our intuitive concept of similarity into Lewis's semantics, we predict the antecedent *if tape A had finitely many holes* to be a counterfactual impossibility. This is clearly an undesirable result.²⁴

This problem also affects Kratzer's semantics of conditionals. Let us consider Kratzer's approach now.

4.3.1 Kratzer's situation-based semantics of conditionals

Kratzer (1989) works in a situation semantics, where situations are parts of possible worlds and propositions defined as sets of situations (as usual, we define that a set of situations p is true at a situation s just in case s is an element of p). Kratzer defines the following relationship of *lumping* between propositions.

- (63) Where p and q are sets of situations, and w a world, p lumps q at w just in case
 a. p is true in w , and
 b. Every situation that is part of w where p is true, q is also true.

Lumping is like entailment, but factive and local: factive in the sense that for a proposition p to lump another at a world w , p must be true at w ; local in the

²⁴Fine (2012b) offers another example where the limit assumption fails and the similarity approach makes the wrong predictions. Fine's scenario involves changes through time (i.e. what worlds are nomically possible), while the tape example is based purely on how things stand at a moment in time. We use the tape example over Fine's to show that problems for the similarity approach remain regardless of nomic possibility.

sense that the lumping relationship is determined with respect to situations in a single world.

Kratzer formulates her truth-conditions for conditionals using premise semantics. She starts with what she calls a *base set* F_w , which is a set of propositions (sets of situations) all of which are true in the world of evaluation w . She imposes some constraints on what counts as an admissible base set (see Kratzer 2012:132–33). These constraints will not play a major role here, since as the reader may verify, they are all satisfied in the examples to come.

Given an admissible base set and a proposition p , Kratzer defines the *crucial set* $F_{w,p}$ as follows.

4.3.1. DEFINITION (The crucial set). For any world w , admissible base set F_w , and proposition p , $F_{w,p}$ is the set of all subsets A of $F_w \cup \{p\}$ satisfying the following conditions:

- (i) A is consistent
- (ii) $p \in A$
- (iii) A is closed under lumping: for all $q \in A$ and $r \in F_w$: if q lumps r in w , then $r \in A$.

4.3.2. DEFINITION (Truth conditions of “would”-counterfactuals). Given a world w and an admissible Base Set F_w , a “would”-counterfactual with antecedent p and consequent q is true in w iff for every set in $F_{w,p}$ there is a superset in $F_{w,p}$ that logically implies q .

In the case of (58), let p be the proposition expressed by *The tape has finitely many holes* and let h_n be the proposition expressed by *Cell n has a hole* where n is a natural number. Propositions such as h_n , which describe a simple fact about the scenario, seem ideally suited to be elements of the base set F_w . So let us assume, naturally enough, that the base set F_w consists of the propositions h_n for every natural number n .

Now take any subset A of $F_w \cup \{p\}$ satisfying conditions (i)–(iii). Since A contains p and is consistent, there must be some number n such that h_n is not in A (otherwise A would entail that there are infinitely many holes on the tape). Let us add h_n to A , resulting in $A^+ = A \cup \{h_n\}$. Then A^+ is also consistent, and is also closed under lumping since no proposition in $F_w \cup \{p\}$ lumps any other proposition in $F_w \cup \{p\}$. Now, we can repeat this operation as many times as we like, adding more and more propositions to A asserting the presence of more and more holes until we reach any finite number of holes – say, one trillion and one. Thus every set in $F_{w,p}$ has a superset in $F_{w,p}$ that logically implies that the tape has more than a trillion holes, so Definition 4.3.2 incorrectly predicts (58), repeated below, to be acceptable.²⁵

²⁵Kratzer (2012:166) also suggests the following notion of exemplification.

- (58) If the tape had finitely many holes, it would still have more than a trillion holes.

4.4 Reconstructing the present proposal within the ordering approach

In the previous section we saw how the clauses for exact verification and falsification interact with our proposal about how sentences raise hypothetical sce-

-
- (i) A possible situation s exemplifies a proposition p iff
- a. p is true at s , and
 - b. p is true at every proper part of s or at no proper part of s .

Exemplification also produces unwanted results when infinity is around. For example, one would like a situation consisting of infinitely many holes on the tape to exemplify the proposition *There are infinitely many holes on the tape*. However, such a situation does not, since it has a proper part (one containing a smaller collection of infinitely many holes) where the proposition expressed by that sentence is true, and also has a proper part (one containing finitely many holes) where the proposition expressed by that sentence is false.

Kratzer (1990, 2002, 2012:171) proposes to get around this problem with the idea that the proposition expressed by, say, *There are infinitely many stars* is true at a situation just in case it contains every star in the world, of which there are infinitely many. This has the unintuitive result that a situation can contain infinitely many stars but still the proposition expressed by *There are infinitely stars* is not true at that situation. Kratzer replies that we can understand such sentences as about all the stars there are, brought out by the German sentence in (ii) in which the noun is topicalized.

- (ii) Sterne gibt es unendlich viele.
Stars are there infinitely many.
 As for the stars, there are infinitely many of them.

There are, however, other sentences where this move is less plausible, such as Fine's *This is moving* (discussed in section 3.7.1), as well as (iii).

- (iii) a. There is at least one infinite collection of stars.
 b. There is a situation containing infinitely many stars.

It is hard to see how (iii) are making a claim about all the stars there are in our world. Kratzer's reasoning here would force us to say that there is a situation containing infinitely many stars where the proposition expressed by *There is a situation containing infinitely many stars* is not true (namely, a situation containing infinitely many but not all stars) – an implausible result.

Now, when we use sentences such as (iii) as a conditional antecedent we have the same problem as above; for example, (iv) are still intuitively unacceptable.

- (iv) a. If there were not one infinite collection of holes on the tape, the tape would still have more than a trillion holes.
 b. If there were no situation containing infinitely many holes, the tape would still have more than a trillion holes.

For further objections against Kratzer's response to this issue, see Deigan (2020:526, n.19).

narios. The result is a theory according to which these alternatives are raised in a systematic way according to the logical form of the sentence. We have seen how the exact relevance clauses for conjunction lead to our proposal invalidating antecedent strengthening, and how the proposed clauses for the quantifiers lead to a semantics of conditionals that makes better predictions than alternative approaches; in particular, those based on similarity (Lewis 1973b), minimal difference (Stalnaker 1968, Pollock 1976), or Kratzer's premise semantics (Kratzer 2012).

While the data in previous section show that our interpretation of conditionals is not based on similarity, minimal difference or Kratzer's premise semantics, they do not show that the purely mathematical architecture of these frameworks is mistaken. One can separate the formal structure of these frameworks from their intuitive concepts. The question then arises whether the semantics of conditionals and causal claims can be expressed using these formal frameworks, abstracted from their original interpretation. One motivation for answering this question is that Kratzer (1977, 2012) has given a semantics for modality in general using an ordering over worlds. One would like a guarantee that the present approach to modality in conditionals and causal claims fits into the Kratzerian big picture.²⁶

In this section we will see that, indeed, under one small modification, it can. The modification is that the order will be relative not only to the world of evaluation but also to the sentence we are asked to imagine true. We will do this by building the present approach inside the ordering and premise semantics to conditionals. The mathematical structure will remain (modulo making the order relative to the sentence in question) but a different intuitive notion will inhabit it, one based on the concepts we have defined here (principally, the notion of an *A*-variant).

Lewis (1981) showed how an ordering could be defined from a set of propositions in the following way.

- (64) For any worlds w', w'' and set of propositions P , define that $w' \leq_P w''$ just in case for all $p \in P$, if $w'' \in p$ then $w' \in p$.

Let g be a function taking a world and returning a set of propositions, what Kratzer calls an *ordering source*. The modal horizon is then taken to be the set of worlds that are closest to the actual world with respect to the order induced

²⁶One small detail is that Kratzer sets time aside in her framework for modality, whereas here the modal horizon is sensitive to the intervention time. Simple examples show the semantics of modals is sensitive to time, such as the following from Fălăuş and Laca (2020).

- (i) a. From next Monday on, Mary will have to wear a uniform at school.
b. Until the beginning of the 90s, students could smoke in class.

See Rullmann and Matthewson (2018) for an account of the semantics of modals that uses the ingredients of Kratzer's framework but makes them time-relative.

by the ordering source.²⁷

$$(65) \quad \min_g(w, a) = \{w' : \text{for no } w'' \in a \text{ is } w'' <_{g(w)} w'\}.$$

In ordering semantics for conditionals (e.g. Stalnaker 1968, Lewis 1973b, Pollock 1976), the relation is stipulated to be reflexive and transitive. Given the definition in (64), however, these properties follow as a matter of logic; from the reflexivity and transitivity of implication, respectively. The definition in (64) therefore makes some progress in stating what determines the order. If we start with an order – as, say, Stalnaker, Lewis and Pollock do – reflexivity and transitivity are imposed ‘top down’, so to speak, while if we start from a set of propositions and assume the definition in (64), reflexivity and transitivity of the order follow from first principles.²⁸

Our question then becomes, what determines the ordering source? Kratzer proposed that in the interpretation of counterfactuals the ordering source is *totally realistic*, meaning that $\bigcap g(w) = \{w\}$ for any world w . That is, the set of propositions in the ordering source specifies the actual world uniquely. Assuming the entry for sufficiency in (66),

$$(66) \quad a \gg_g c \text{ iff for all } w' \in \min_g(w, a), \text{ if } w' \in a \text{ then } w' \in c.$$

total realism validates the rule:

$$(67) \quad \text{If } g \text{ is totally realistic, } a \wedge c \text{ entails } a \gg_g c.$$

Given the data in (3)–(9), this is obviously a pattern we wish to avoid. The truth of both A and C does not imply that A is sufficient for C . For example, the fact that Ali was born in Europe and has an Irish passport does not imply that him being born in Europe was sufficient for him to have an Irish passport. The fact that one is over twelve years old and can order wine does not imply that being over twelve years old is sufficient to order wine.

Kratzer also defines a weaker constraint: an ordering source is *realistic* just in case for every world w , every proposition in $g(w)$ is true in w ; that is, $w \in \bigcap g(w)$. Realism validates the following desirable rule.

$$(68) \quad \text{If } g \text{ is realistic, } a \gg_g c \text{ and } a \text{ together entail } c.$$

²⁷The strict version of the order is defined as usual: $x < y$ just in case $x \leq y$ but not $y \leq x$.

²⁸It is sometimes said that Lewis (1981) showed that premise semantics (Veltman 1976, Kratzer 1981a,b) and ordering semantics are “notional variants” of one another. Given the remarks above, one could say that premise semantics is in some sense deeper than ordering semantics, since it derives from general facts about implication properties that the ordering approach stipulates. Pushing this thought further, we can think of premise semantics as giving an analysis of the concept of similarity itself, which could be applied not only to worlds but to any entities, the premises being those features we take into account when judging similarity. This would account for why comparative similarity is reflexive and transitive in the first place.

In addition to realism, what other constraints should the ordering source satisfy? One constraint, noticed by Veltman (1976:266–67) and Kratzer (1981a), is that if $g(w)$ contains *every* proposition true at w , we get the wrong results. To illustrate, recall the switches in Figure 3.1. Consider the following propositions.

- (69) b Switch B is up.
 d The switches are in a different position.

In the actual scenario, A is down and B up, so b and d are both true. Now, when we imagine switch A up, we intuitively imagine both switches being up: b stays true while d goes from true to false. This shows that the ordering source should contain b but not d . It should contain b since we want worlds where switch B has its actual position to be closer than worlds where it has a different position. Switch B should not change gratuitously. And the ordering source should not contain d , since if it did, by (64) and (65), the scenario where A is up and B down would sneak into the modal horizon. Now, our question is: what general principle accounts for why b is in the ordering source but d is not?

A first thought is to look to entailment relationships. We are given the conditional antecedent *if switch A were up*. Let a be the proposition expressed by *Switch A is up*. Do b and d differ in their logical relationship to a ? It turns out they do not. To see this, let us say two propositions are *orthogonal* just in case (i) p does not entail q , (ii) p does not entail $\neg q$, (iii) $\neg p$ does not entail q , and (iv) $\neg p$ does not entail $\neg q$.²⁹ It turns out that a and b are orthogonal just in case a and d are. So b and d bear the same entailment relationship to the antecedent. If we want to distinguish b and d , we will have to look elsewhere.

While we have focused on a specific example, the problem is fully general. Let a be any false proposition that we are asked to imagine true, and let b be any true proposition whatsoever that intuitively stays true when we imagine a true. Consider the proposition $d = (a \leftrightarrow \neg b)$. Then b and d are both true, and b should be in the ordering source while d should not (if d were in the ordering source, the modal horizon when imagining a true would contain a world where b is false). But logical relationships alone do not see the difference, for it is a fact of propositional logic that a and b are orthogonal just in case a and $a \leftrightarrow \neg b$ are.

Here is one suggestion of what is going on. When we are given a scenario where a is false, and asked to imagine a true in that scenario, a proposition should not matter for similarity if its truth in part ‘depends’ on the fact that a is false. Changing a proposition can also change the propositions that depend on it. For example, the fact that the switches are in a different position depends on the actual positions of the switches; in this case, the fact that A is down and B

²⁹This definition of orthogonality may seem asymmetric, but in fact it is symmetric: if p and q are orthogonal then q and p are orthogonal. This holds due to contraposition; for example, notice that p does not entail q just in case $\neg q$ does not entail $\neg p$. The use of the term *orthogonal* here comes from Lewis (1988).

is up. So the truth of d depends in part on the fact that A is down. In contrast, the fact that B is up intuitively does not depend on the fact that A is down. So b matters for similarity while d does not.

This seems to be a promising, informal start. With it we can already explain why logical relationships cannot distinguish between the facts that do and do not matter for similarity. For this notion of dependence between propositions is sensitive to the particular facts of the actual scenario. For example, in the actual scenario, where A is down and B up, the fact that the switches are in a different position depends on the fact that A is down and B is up. In a scenario where the positions are reversed – A is up and B is down – that fact that the switches are in a different position would instead depend on a different fact; namely, that A is up and B is down. It follows that logical relationships such as entailment cannot capture the notion of dependence we are after. For logical relationships are global: they are defined over all of logical space, blind to the particular facts of the actual scenario.

What we would like to do now is make these informal remarks precise. One person in particular who has made great progress on this front is Angelika Kratzer.

4.4.1 Sufficiency in Kratzer’s semantics of conditionals

Recall Kratzer’s semantics of conditionals, introduced in section 4.3.1.

- (70) Where p and q are sets of situations, and w a world, p lumps q at w just in case
- a. p is true in w , and
 - b. Every situation that is part of w where p is true, q is also true.

Given an admissible base set and a proposition p , Kratzer defines the *crucial set* $F_{w,p}$ as follows.

4.4.1. DEFINITION (The crucial set). For any world w , admissible base set F_w , and proposition p , $F_{w,p}$ is the set of all subsets A of $F_w \cup \{p\}$ satisfying the following conditions:

- (i) A is consistent
- (ii) $p \in A$
- (iii) A is closed under lumping: for all $q \in A$ and $r \in F_w$: if q lumps r in w , then $r \in A$.

4.4.2. DEFINITION (Truth conditions of “would”-counterfactuals). Given a world w and an admissible Base Set F_w , a “would”-counterfactual with antecedent p and consequent q is true in w iff for every set in $F_{w,p}$ there is a superset in $F_{w,p}$ that logically implies q .

This framework can account for why, given the antecedent *if switch A were up*, we can keep b (switch B is up) fixed but allow d (the switches are in a different position) to vary, and thus why we interpret (71) to be true.

(71) If switch A were up, both switches would be up.

It is natural to assume that the base set F_w contains the fact that switch B is up. Then, where a is the proposition that switch A is up, every set in $F_{w,a}$ contains both a and b . So $F_{w,a}$ logically implies that both switches are up and (71) is correctly predicted to be true. In contrast, constraints (i)–(iii) prevent d from being in the base set. They even imply that, for any base set F_w whatsoever, d is not in any subset of the crucial set $F_{w,a}$. To see this, let A be any subset of $F_{w,a}$ and suppose for reductio that A contained d . Note that d lumps $\neg a$, since the switches actually are in a different position, and every situation that is part of the actual world where the switches are in a different position is also situation where switch A is down. Since A contains d and d lumps $\neg a$, by (iii), A also contains $\neg a$. But by (ii), A contains a , so A is inconsistent, contradicting (i). Hence, as desired, A does not contain d .

Here is another example to help appreciate how Kratzer's framework makes the right predictions for conditionals. Suppose we are looking at a can of ultramarine paint, and consider (72).

(72) If the paint were not ultramarine, it would still be blue.

There may be some background reasons to keep fixed the fact that the paint is blue; we find ourselves in a factory that only makes blue paint. But without such specific contextual constraints, there is no reason to imagine that if the paint were not ultramarine, it would still be blue. Kratzer's approach correctly accounts for this, in the same way it makes the right predictions for (71). The key to this result in Kratzer's framework is (73).

- (73) The proposition that the paint is blue lumps the proposition that it is ultramarine at w :
- a. the paint is blue at w , and
 - b. every situation that is part of w where the paint is blue is a situation where it is ultramarine.

Let $\neg u$ be the proposition that the paint is not ultramarine. Then for any base set F_w whatsoever, no subset A of the crucial set $F_{w,\neg u}$ contains the proposition that the paint is blue. If A did contain the proposition that the paint is blue, by (73) and closure under lumping, A would also contain u , but as A contains $\neg u$, A would be inconsistent. Given the fact that blue lumps ultramarine, when we imagine the paint being a colour other than ultramarine, Kratzer's approach correctly predicts that we do not fix the fact that it is blue.

This illustrates how Kratzer's framework makes the right predictions concerning what we fix and what we allow to vary when interpreting conditionals such as (71) and (72). Now, as discussed in section 3.2, we would like a framework that represents the imaginative faculty underlying our interpretation of causatives as well as conditionals. It is natural, therefore, to wonder whether Kratzer's approach can handle causatives as well.

It turns out it cannot, since it cannot capture sufficiency. Recall the sufficiency violations in (3)–(9). Take for example (7), repeated below.

- (7) a. Yves bought paint C because it is blue.
 b. Yves bought paint C because it is ultramarine.

When we imagine paint C being blue, we are not forced to imagine it having the particular shade of blue it actually has. We can imagine it azure, baby blue, sky blue, and so on. However, if we try to apply Kratzer's framework for conditionals to analyse sufficiency, we are forced to fix the paint to its actual shade of blue. That is, given that the paint is ultramarine, Kratzer predicts (74) to be true.

- (74) If the paint were blue, it would be ultramarine.

This is due to how Kratzer's framework raises hypothetical scenarios (independently of any selection function in the semantics of conditionals). To see this, let b be the proposition that the paint is blue. Let F_w be any base set, and A any subset of $F_{w,b}$. By (ii), A contains b . But then since b lumps u , by closure under lumping A contains u . So A has a superset that logically implies u (indeed, *every* superset of A logically implies u). The upshot is that Kratzer's framework makes the right predictions for conditionals such as (72) at the expense of being unable to capture sufficiency. If we try to apply the framework to the conditionals in (7), we wrongly predict the equivalence of (7a) and (7b). Of course, (7) is not unique in this regard: the same can be said of the other contrasts observed in (3)–(9).

At this point, it is natural to ask just how deep this inability to capture sufficiency runs in Kratzer's framework. To make the right predictions for both conditionals and causatives, do we need a new framework altogether, or is a superficial modification enough? It turns out that, as long as the framework is based on lumping, it cannot capture sufficiency. The problem is that lumping is not fine-grained enough to give us all the distinctions we need to capture sufficiency, as we will see now.

4.4.2 From lumping to overlap

Consider the following small variation on the ultramarine paint example. Yves wants *two* tins of ultramarine paint. He enters a paint shop selling tins A–H (Figure 4.13). Four of these are blue (B, C, E and G), of which two are ultramarine (C and E). He does not buy cans B or G, as they are the wrong shade of blue.

He buys cans C and E.

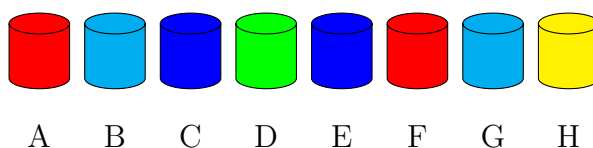


Figure 4.13

Consider (75) in this context.

(75) Yves bought tins C and E because they are both blue.

(75) is intuitively unacceptable. There is a contrast in acceptability between (75) and (76).

(76) Yves bought tins C and E because they are both ultramarine.

When interpreting (75), we do not fix the fact that the tins are ultramarine. At the same time, there are many facts we do fix, such as the fact that Yves wants ultramarine paint, the fact that C and E are for sale, the colours of the other tins, and countless other facts. The challenge, of course, is to distinguish the facts we fix from those we allow to vary when interpreting (75). Is lumping up to the task? It turns out it is not. Consider the propositions in (77).

- (77)
- a. C is blue and E is blue.
 - b. C is ultramarine and D is green.
 - c. E is ultramarine and D is green.

When we interpret (75), we do not hold fixed (77b) and (77c). For they together conspire to keep the cans ultramarine. But then there would be no contrast between (75) and (76). However, (77a) does not lump (77b): in the situation containing only tins C and E, it is true that C is blue and E is blue, but not that C is ultramarine and D is green. Similarly, (77a) does not lump (77c). A drop of irrelevant information (e.g. that D is green) breaks the lumping relation. The result is that the cause – that C is blue and E is blue – bears the same lumping relations to propositions we do not hold fixed, such as (77b) and (77c), as it those we do hold fixed (e.g. the fact that Yves wants ultramarine paint, that tins C and E are for sale, and so on). Lumping is not fine-grained enough to distinguish the facts we fix from those we allow to vary.

Our question now is what more fine-grained notion we could use instead of lumping. Towards an answer to this question, notice that there is a sense in which (77a) ‘overlaps’ (77b) and (77c). They each have a ‘part’, so to speak, in common: (77a) and (77b) both involve the colour of tin C; (77a) and (77c) both involve the colour of tin E. Perhaps, then, when we evaluate whether the truth of one proposition *a* is sufficient for the truth of another proposition *c*, a proposition

p should not matter for similarity just in case it in some sense ‘overlaps’ a at the world of evaluation. That is, given a world w and proposition a , the ordering source $OS(w, a)$ should be the set of propositions that are true at w and do not ‘overlap’ a at w .

$$(78) \quad OS(w, a) := \{p : p \text{ is true at } w \text{ and } p \text{ does not overlap } a \text{ at } w\}.$$

4.4.3 Definitions of overlap

But what does it mean for two propositions to overlap at a world? Our guiding example is that *C is blue and E is blue* should overlap *C is ultramarine and D is green* and *E is ultramarine and D is green*. One initially plausible attempt is:

- (79) **Definition of overlap (first attempt).** Propositions p and q overlap at world w just in case p and q are both true in w , and there is a proposition r such that
- a. p entails r and q entails r , and
 - b. w contains a situation where r is true.

This gives the right result for our guiding example. According to it, *C is blue and E is blue* overlaps *C is ultramarine and D is green* since, taking r to be *C is blue*, the world of evaluation w contains a situation where C is blue, and

$$\begin{array}{ll} C \text{ is blue and } E \text{ is blue} & \text{entails } C \text{ is blue} \\ C \text{ is ultramarine and } D \text{ is green} & \text{entails } C \text{ is blue.} \end{array}$$

Similarly, taking r to be *E is blue*, we find that *C is blue and E is blue* overlaps *E is ultramarine and D is green*.

However, our first attempt is far too permissive. According to it, truth entails overlap: if two propositions are true at a world then they overlap at that world. This is because we can simply take r to be $p \vee q$. For whenever p and q are true at a world, p and q entail $p \vee q$, and w contains a situation (namely, itself) where $p \vee q$ is true. Then given (78), when a is true at w (as it is when evaluating sufficiency), according to our first attempt at defining overlap, $OS(w, a)$ is empty. And with an empty ordering source, sufficiency amounts to entailment. To see this, recall how we analysed sufficiency using the modal horizon, the modal horizon using the order, and the order using the ordering source, with the definitions below (suitably adapted to make the ordering source sensitive to the antecedent), for any world w and propositions (sets of worlds) a and c .

$$\begin{aligned} w'' \leq_{OS(w, a)} w' & \Leftrightarrow \text{for all } p \in OS(w, a), \text{ if } w' \in p \text{ then } w'' \in p \\ w'' <_{OS(w, a)} w' & \Leftrightarrow w'' \leq_{OS(w, a)} w' \text{ and } w' \not\leq_{OS(w, a)} w'' \\ \min(w, a) & = \{w' : \text{for no } w'' \in a \text{ is } w'' <_{OS(w, a)} w'\} \end{aligned}$$

Then if $OS(w, a)$ is empty, the order is trivial: vacuously, for all worlds w' , w'' whatsoever, $w' \leq_{OS(w, a)} w''$, so it is not the case that $w'' <_{OS(w, a)} w'$. Then the modal horizon $mh(w, a)$ is the set of all worlds, and sufficiency collapses to entailment. This does not reflect the systematic nature of the imagination, whereby we are able to carry over some facts from the actual world into the hypothetical scenarios we imagine.

Does replacing entailment with lumping help? A second plausible attempt is:

- (80) **Definition of overlap (second attempt).** Propositions p and q overlap at world w just in case there is a proposition r such that p lumps r at w and q lumps r at w .

This also gives the right result for our guiding example. According to our second attempt, *C is blue and E is blue* overlaps *C is ultramarine and D is green* since, again taking r to be *C is blue*, we have:

$$\begin{array}{ll} C \text{ is blue and } E \text{ is blue} & \text{lumps } C \text{ is blue} \\ C \text{ is ultramarine and } D \text{ is green} & \text{lumps } C \text{ is blue.} \end{array}$$

Similarly, taking r to be *E is blue*, we find that *C is blue and E is blue* overlaps *E is ultramarine and D is green*, as desired. However, our second attempt fails for the same reason as the first. According to it, if two propositions are true at a world, they overlap at that world. For if p and q are both true at w , then p and q each lump $p \vee q$ at w . So our second attempt also collapses sufficiency to entailment.

A third, more hopeful attempt is:

- (81) **Definition of overlap (third attempt).** Propositions p and q overlap at a world w just in case there is a state that is part of w , part of a state where p is true, and part of a state where q is true.
If p does not overlap q at w , we say p and q are disjoint at w .

On this attempt too, truth entails overlap. For we can simply take s and t to both be w itself. Then if p and q are both true at w , w contains a situation (namely, itself) that is part of a situation where p is true and part of a situation where q is true, so p overlaps q at w . Yet again, our attempt collapse sufficiency to entailment.

While this third attempt also fails, it fails better than the previous attempts, since it gives us a clear view of why it fails. It fails because it allows states that are ‘too big’ to witness an overlap between propositions, states with parts that are irrelevant to the truth of the propositions in question. An obvious solution, inspired by Fine (2017b), is to require that s and t be *exactly* relevant to p and q , respectively. In terms of the present framework, we require s and t to be in the respective foregrounds of p and q .

- (82) **Definition of overlap (final).** Propositions p and q overlap at a world w just in case there is a state that is part of w , part of a state in the foreground of p and part of a state in the foreground of q .

This raises the question what it means for a state to be in the foreground of a proposition. While we have not provided a complete answer to that here, in the examples we have considered so far I believe the notion is perfectly clear. It is natural to suppose that there is a state c – call it the state of C’s colour – in the foreground of *paint C is ultramarine* and *paint C is blue*, and similarly a the state e of E’s colour. There is also their fusion $c \sqcup e$ in the foreground of *C is blue and E is blue*. Then our definition of overlap in (82) gives the right results.

- (83) *C is blue and E is blue* overlaps *C is ultramarine and D is green* since
- a. c is part of w , $c \sqcup e$ and $c \sqcup d$.
 - b. $c \sqcup e$ is in the foreground of *C is blue and E is blue*, and
 - c. $c \sqcup d$ is in the foreground of *C is ultramarine and D is green*.

In our running example of the switches, the foregrounds are also intuitively clear. There is the state of switch A being down in the foreground of *switch A is down*, and the state of switch B being up is in the foreground of *switch B is up*. These states do not overlap. Assuming these are the only states in the foreground of the respective statements where these sentences are true, we get the desired result from (82) that *switch A is down* and *switch B is up* do not overlap, so when we imagine a change to switch A, we keep the fact that switch B is up.

4.4.4 Proving the equivalence of the ordering semantics and the present approach

In this section we prove that, under some auxiliary assumptions, the present approach can be expressed via an ordering over worlds, where the order is relative to the world and sentence of evaluation. We will be deliberately vague on what counts as a proposition, since different frameworks use the term in different ways and our present goal is to translate between frameworks. We take the foreground to assign to each proposition a set of states, and take truth at a world as a relation between propositions and worlds.

- (84) a. For every world w and proposition p , if p is true at w then w contains a state s such that
- (i) s is in the foreground of p ,
 - (ii) s settles p : p is true at every world containing s or p is false at every world containing s .
- b. For every state s , there is a proposition *actual*(s) such that
- (i) for any world w , *actual*(s) is true at w iff s is part of w ;
 - (ii) every state in the foreground of *actual*(s) is part of s .

Recall that a world w' is $\leq_{w,a}$ -minimal just in case for no world w'' is $w'' <_{w,a} w'$.

4.4.3. PROPOSITION. *Given the assumptions in (84), for any worlds w, w' and proposition a ,*

$$w' \text{ is } \leq_{w,a}\text{-minimal} \quad \text{if and only if} \quad w' \text{ is an } a\text{-variant of } w.$$

PROOF. (\Rightarrow) Suppose w' is $\leq_{w,a}$ -minimal. Then $w \not\leq_{w,a} w'$, i.e. $w \not\leq_{w,a} w'$ or $w' \leq_{w,a} w$. Now, $w \leq_{w,a} w'$. So $w' \leq_{w,a} w$, i.e. (*) for all $p \in OS(w, a)$, if $w \in p$ then $w' \in p$. We show that w' is an a -variant of w : every part s of w that does not overlap a state in the foreground of a is part of w' . Pick any such s . By (84b), there is a proposition $actual(s)$. We show that $actual(s) \in OS(w, a)$. Since s is part of w , by (84b-i), $actual(s)$ is true at w . Also $actual(s)$ does not overlap a at w . For suppose for reductio that it did. Then by (82), there is a state u that is part of a state s' in the foreground of $actual(s)$ and a state in the foreground of a . By (84b-ii), s' is part of s , so u is part of s , contradicting the fact that s does not overlap a state in the foreground of a . So $actual(s)$ is true at w and does not overlap a at w : $actual(s) \in OS(w, a)$. Then by (*), $actual(s)$ is true at w' , so by (84b-i), s is part of w' .

(\Leftarrow) Pick any a -variant w' of w and $p \in OS(w, a)$. We show that $w' \in p$. Since $p \in OS(w, a)$, $w \in p$. Then by (84a), w contains a state s in the foreground of p that settles p . We show that s does not overlap any state in the foreground of a . For suppose it did. Then some state t is part of a state in the foreground of p and a state in the foreground of a . And as t is part of s and s is part of w , t is part of w . But then p and a overlap at w , contradicting the fact that $p \in OS(w, a)$. So s is part of w and does not overlap any state in the foreground of a . Then as w' is an a -variant of w , s is part of w' . Since s settles p , s is part of w and p is true at w , p must be true at every world containing s . Hence p is true at w' . Then for any $p \in OS(w, a)$ and $w'' \in a$, $w' \leq_{w,a} w''$, and so $w'' \not\leq_{w,a} w'$. Therefore w' is $\leq_{w,a}$ -minimal. \square

Let $min(\leq_{w,a})$ be the set of $\leq_{w,a}$ -minimal worlds, and $a\text{-variants}(w)$ be the set of a -variants of w . Then given the assumptions in (84), by Fact 4.4.3 we have the following equivalences, for any world w and propositions a and c ,

$$\begin{aligned} min(\leq_{w,a}) &= a\text{-variants}(w) \\ min(\leq_{w,a}) \cap a \subseteq c &\Leftrightarrow a\text{-variants}(w) \cap a \subseteq c \end{aligned}$$

which shows that the present approach to sufficiency can be expressed in terms of the ordering approach. The present approach is simpler than the ordering approach, since it operates on states directly; indeed, one can calculate the a -variants of a world w by considering a single state: $w - a\text{-part}(w)$ (defined in section 3.6.1). In contrast, the ordering approach takes into account unfathomably many propositions. Proposition 4.4.3 shows that the additional architecture of the ordering approach, while baroque, is compatible with our approach.

Nonetheless, it is interesting to see that we can express the present proposal in terms of the ordering approach. For we saw some evidence in section 4.3 that the semantics of counterfactuals is not based on our intuitive concept of similarity, nor based on the notions of lumping and consistency as used in Kratzer’s approach. Moreover, in this section we have seen that these approaches cannot capture how we interpret causal claims, since they cannot capture sufficiency. Then the fact that we can express the present approach in terms of orderings shows that the difficulties with the similarity approach and premise semantics lie not with their mathematical framework but with the intuitive notions on which they are based.

4.5 Conclusion

In this chapter we extended our analysis of sufficiency and *would*-conditionals from Chapter 3 to logically complex sentences. This allowed us to compare our proposal with existing accounts of the semantics of conditionals (section 4.3). We saw cases where the present approach gives more accurate predictions than existing theories that use similarity orders and Kratzer’s premise semantics, and we proved that the present approach can be phrased using the formal architecture of these accounts by replacing their primitive notions with the idea of varying a state (section 4.4). Finally, we considered a surprising rule that our approach validates (or-to-and, from section 4.2.2), and proposed a way to account for the data while preserving the rule’s semantic validity.

Chapter 5

An analysis of production

5.1 Introduction

In chapter 2 we proposed a semantics of *cause* and *because* in terms of two relations: sufficiency and production. Specifically, where \gg expresses sufficiency, we proposed that for any sentences C and C, E *because* C is true if and only if

$$C \wedge (C \gg (C \text{ produce } E)) \wedge \neg(\neg C \gg (\neg C \text{ produce } E))$$

is true; in other words, if and only if C is true, and C sufficient for it to produce E but $\neg C$ is not. Similarly, for any noun c and *to*-infinitive e , we proposed that c *cause* e is true just in case the above formula holds for the sentences expressed by c and e . Chapter 3 gave a formal analysis of sufficiency. The goal of the present chapter is to round off our analysis with a formal analysis of production.

The use of production in the analysis in causal claims goes back to Hall (2004), who contrasted production with counterfactual dependence. Hall argued that “there are two *kinds* of causation, two different ways in which one event can be a cause of another”, which he viewed as irreconcilable in virtue of their different properties: production is assumed to be transitive, spacio-temporally local, and in some sense intrinsic to the events involved, while counterfactual dependence is not. Hall (2007) later abandoned a ‘two concepts’ view,¹ but the distinction between production and dependence remains an important milestone in work on causation. It appears under different guises today; for example, Gerstenberg et al. (2021) distinguish between *whether causation* and *how causation*, with *whether* causation similar to dependence and *how* causation to production. Production might also, perhaps, be the notion that causal process theories of causation (such as Salmon 1984, Dowe 2000) and force-dynamics approaches (such as Talmy 1988, Wolff 2007, Copley and Harley 2015) aim to describe.

¹Hall (2007) does not directly argue against a ‘two concepts’ view, but merely proposes a univocal analysis in its place. For a direct argument against a ‘two concepts’ view, see Corkum (2022).

Sander Beckers, in his PhD thesis (Beckers 2016) and subsequent paper (Beckers and Vennekens 2018) gives a semantics of the expression *is an actual cause of* that blends dependence and production. Beckers defines these notions in terms of structural causal models. For simplicity we will not state his formal conditions here. Loosely put, Beckers (2016:95) proposes:

- (1) C is an actual cause of E just in case
 - a. C and E are true;
 - b. C produced E ; and
 - c. if C had not been true, $\neg C$ would not have produced E .

Let us call (1c) *Beckers' difference-making condition*. Its blend of dependence and production represents, I believe, a remarkable breakthrough in the analysis of causation, and in particular, our understanding of overdetermination – cases of causation without counterfactual dependence. Our own difference-making condition above, stating that $\neg C$ is not sufficient for it to produce E , owes an obvious debt to Beckers.

5.2 Motivating production

Before getting into formal details, let us briefly motivate Beckers' difference making condition. Recall the Billy and Suzy from Hall (2004:235).

Suzy and Billy, expert rock-throwers, are engaged in a competition to see who can shatter a target bottle first. They both pick up rocks and throw them at the bottle, but Suzy throws hers before Billy. Consequently Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle if Suzy's had not occurred, so the shattering is overdetermined.

- (2)
 - a. The bottle broke because Suzy threw her rock at it.
 - b. Suzy throwing her rock at the bottle caused it to break.
- (3)
 - a. The bottle broke because Billy threw his rock at it.
 - b. Billy throwing his rock at the bottle caused it to break.

(2) are intuitively acceptable, while (3) are not. Beckers (2016) accounts for this by giving an analysis of production where Suzy produced the bottle to break, and if Suzy hadn't thrown, Suzy *not* throwing would not have produced the bottle to break (instead, Billy throwing would have produced the bottle to break). On the other hand, Billy did not produce the bottle to break. While Billy fails the production condition, he satisfies the difference-making condition: if Billy hadn't thrown, him not throwing would not have produced the bottle to break. Thus on Beckers' proposal it is the production condition that accounts for the contrast between (2) and (3).

Compare this with the switching following scenario from Hall (2000:205).

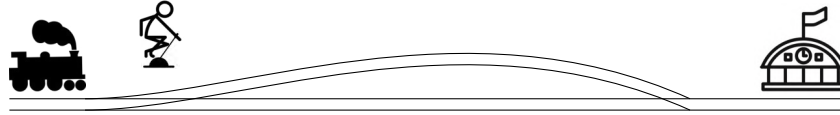


Figure 5.1: Hall's switching scenario

An engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the bottom, instead of top track. Since the tracks reconverge up ahead, the train arrives at its destination all the same.

Consider (4) in this context.

- (4)
 - a. The train reached the station because the engineer pulled the lever.
 - b. The engineer pulling the lever caused the train to reach the station.

(4) are intuitively unacceptable. On our semantics for *cause* and *because*, there are in principle two ways to account for this:

Option 1: The positive condition fails. Pulling the lever was not sufficient for pulling the lever to produce the train to reach the station.

Option 2: The negative condition fails. Not pulling the lever was sufficient for not pulling the lever to produce the train to reach the station.

To answer which condition fails, let us imagine a slightly different case: instead of the tracks converging, they diverge (Figure 5.2). The engineer pulled the lever sending the train toward the station. If she hadn't pulled the lever, the train would have missed the station.

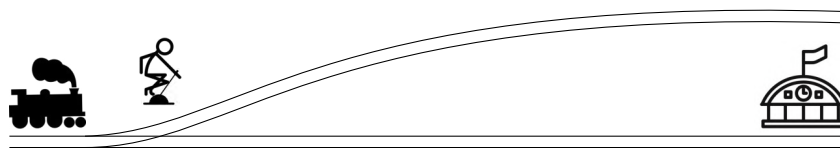


Figure 5.2: Divergence context.

In the divergence context (4) are perfectly fine. To predict this, both the positive and negative conditions must hold: pulling the lever is sufficient for that to produce the train to reach the station, and not pulling the lever is not sufficient for that to produce the train to reach the station.

It is plausible to assume that the convergence and divergence contexts are symmetric with respect to production; that is, pulling the lever produced the

train to reach the station in the convergence context just in case it did so in the divergence context. This is plausible in light of Hall's view that production is spatio-temporally local and intrinsic to the process from cause to effect. In fact, the two contexts agree on what actually happened: the engineer pulled the lever, the train took the bottom track and reached the station. On Beckers' proposal it is therefore the difference making condition that accounts for the difference in (4)'s acceptability in the convergence and divergence contexts: it fails in the convergence context and holds in the divergence context.

Table 5.1 summarises the role of production and difference-making in the scenarios we discussed in this section.

	Suzy	Billy	Lever (divergence)	Lever (convergence)
<i>Intuitively a cause</i>	✓	✗	✓	✗
<i>Production</i>	✓	✗	✓	✓
<i>Difference-making</i>	✓	✓	✓	✗

Table 5.1: The role of production and difference-making in Beckers (2016).

5.2.1 Beckers' analysis of production

Becker's provides an analysis of production in terms of structural causal models, restricting attention to literals (a literal is an atomic sentence or a negated atomic). Informally, Beckers' defines production as follows (for formal details see Beckers 2016: chapter 6).

- (5) **Beckers' analysis of production.** For any literals C and E ,
 C produced E just in case there is a chain of literals C_1, \dots, C_n such that
- a. The chain begins with C and ends with E ($C = C_1$ and $C_n = E$);
 - b. For each C_i and C_{i+1} on the chain, there is a set of literals L where
 - (i) Each sentence in L is true;
 - (ii) Each sentence in L became true before (or simultaneous with) C_{i+1} becoming true;
 - (iii) L is sufficient for C_{i+1} but $L \setminus \{C_i\}$ is not sufficient for C_{i+1} .

In other words, Beckers' analyses production as a chain of NESS tests that respects the order of time.² Each condition on the chain must be a necessary

²For an analysis of causation using the NESS test, see Wright (1985, 2011). The NESS test shares many similarities with Mackie's INUS condition, stating that a cause is "an insufficient but non-redundant part of a condition which is itself unnecessary but sufficient for the result (Mackie 1974:64). The NESS test drops Mackie's condition that the cause be insufficient for the effect. This is an improvement, since causes can be sufficient for their effects by themselves, as arguably shown in the following naturally-occurring examples (discussed in section 2.7 above).

element of a sufficient set for the next, and the sufficient set in question must not contain conditions that become true after the next condition on the chain has become true.

5.2.2 Sufficiency in Beckers' account of production

A key difference between Beckers' definition of sufficiency and our own is that Beckers' requires each event on the chain to be sufficient for the next with respect to a given set of background conditions. However, it turns out that C can produce E without each event on the production chain being sufficient for the next. To illustrate, consider the following variation of the divergence context.

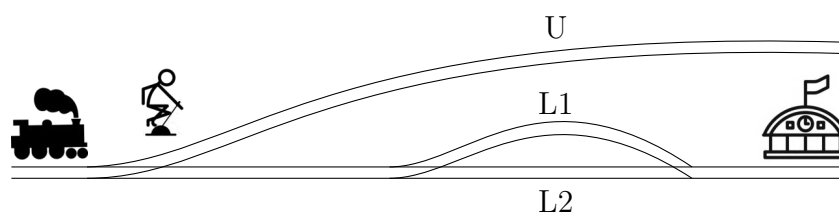


Figure 5.3: Divergence context with some randomness.

The engineer controls whether the train takes the upper track or not. The lower track itself contains a fork, and the train driver decides at random whether to take L1 or L2. The train was originally set to take the upper track, but the engineer pulled the lever, making it take L1 or L2. The driver then decided at random which one of the lower tracks to take; say, it took L1. The train reached the station.

Consider (4), repeated below, in this context.

- (4)
 - a. The train reached the station because the engineer pulled the lever.
 - b. The engineer pulling the lever caused the train to reach the station.

In this context (4) are fine. To predict this on the present account, we need to say that the engineer pulling the lever produced the train to reach the station. The scenario is designed so that the engineer pulling the lever is not sufficient for the train to take track L1, and not sufficient for it to take track L2.

While we modify Beckers' definition to not require sufficiency, we nonetheless take Beckers' analysis of production as a starting point for our own analysis. Let us give our analysis of production now.

- (i)
 - a. Computers do an awful lot of deliberation, and yet their every decision is wholly caused by the state of the universe plus the laws of nature. [\[Source\]](#)
 - b. If you keep asking "why" questions about what happens in the universe, you ultimately reach the answer "because of the state of the universe and the laws of nature." [\[Source\]](#)

5.3 Production in terms of sufficiency

Consider a row of dominoes – a paradigm case of causation if ever there was one. Domino A knocks over domino B, which knocks over domino C. In this context the following sentences are perfectly fine.

- (6) a. Domino A falling caused domino C to fall.
 b. Domino C fell because domino A fell.

Given that *cause* and *because* require the cause to produce the effect, to account for these sentences we have to say that domino A falling produced domino C to fall. There are many to analyse such a chain; for example, as a transmission of a mark (Salmon 1984), conserved quantity (Salmon 1994) or force (Talmy 1988, Wolff 2007, Copley and Harley 2015). We will not consider arguments against these approaches here. We will instead analyse the chain as a chain of *counterfactual dependence*: Domino C falling counterfactually depended on domino B falling, which counterfactually depended on domino A falling.

- (7) **The idea.** For any sentences C and E , C produces E just in case there is a chain of counterfactual dependence from C to E .

To formally analyse this idea (7) we must answer the following four questions.

1. What is a chain, mathematically speaking?
2. What is the chain made up of? What are its relata?
3. What notion of “counterfactual dependence” is used here?
4. What do “from C ” and “to E ” mean?

5.3.1 What is a chain?

Recall from section 3.6.3 the modelling framework we assume, where moments are maximal situations with respect to parthood, representing how things stand at a moment in time, and worlds are linear orders of moments.

Given a world w , let us say that an *interval* I of $w = (M, \preceq)$ is a convex set of moments of w , i.e. $I \subseteq M$ for all $t, t' \in I$ and moment $t'' \in M$ such that $t \preceq t'' \preceq t'$, we have $t'' \in I$. The requirement that I be convex will prevent chains from having any gaps (having no gaps, after all, seems to be part of what it means to count as a ‘chain’). A chain will be made up of situations, one for each moment in a time interval. We will also require that the interval is nonempty and has a minimal and maximal element. This implies that the chain has a first and last element.

5.3.1. DEFINITION (Chain). For any world w , a chain of w is a set of situations $\{s_t\}_{t \in I}$ such that I is a nonempty interval of w with a minimal and maximal element, and for all $t \in I$, s_t is part of t .

Note that this definition allows a chain to consist of a single element. We may rule out these cases by defining a notion of *proper* chain.

5.3.2. DEFINITION (Proper chain). A chain $\{s_t\}_{t \in I}$ is *proper* just in case I has at least two elements.

The distinction between proper improper chains is useful when accounting for differences in meaning between *cause* and *because*. Let us turn to those differences now.

5.4 Proper and improper chains with *cause* and *because*

So far we have been silent about any differences between *cause* and *because*. One difference between them is that *because* has reflexive uses while *cause* does not. Compare:

- (8) a. Me doing it caused me to do it.
b. I did it because I did it.

(8a), if it is acceptable at all, can only mean that doing it once caused me to do it a second time. The cause and effect cannot be the same. Not so with (8b). Here are some naturally-occurring examples of reflexive *because*.³

- (9) a. I just did it because I did it.
b. Atkinson's Sept. 30 statement defending his decision to deem the complaint "credible" amounts to: "I did it, because I did it." He never gave a reason.
c. To this day, I don't know why that anxiety erupted at that time. It happened because it happened.
d. Everything I've done has felt very natural, and it's happened because it's happened.
e. It happened because it happened.
f. But it is the way it is just because it is the way it is.

This suggests that speakers can take reflexive *because* claims can be true. Now, why would a speaker ever assert a reflexive *because* claim? One reason could be to express that speaker is not in a position to assert any more informative answers to the same question. This is expected if, whenever E is true, E because E is true too. For then E because C entails E because E for any C whatsoever, making E because E a least informative member of the set of alternatives $\{E$ because C :

³Sources: (9a) [Spokesman.com](#), 25 September 1996; (9b) [New York Post](#), 30 December 2019; (9c) [Adventurouskate.com](#); (9d) Johnny Depp, [Brainyquote.com](#); (9e) Jack Straw, [UK Parliament](#), 24 June 2003; (9f) [StackExchange](#), 18 March 2018.

C is a sentence}. Reflexive *because* is a good option when you want to give a true answer to a *Why?* question but do not have anything informative to say.

If we allow improper chains in the notion of production used by *because*, we indeed predict that E entails E *because* E . Recall our intuitive idea that C produces E just in case there is a chain of counterfactual dependence from C to E . Every sentence trivially counterfactually depends on itself, so E produce E reduces to E , and therefore the whole semantics of E *because* E reduces from

$$E \wedge (E \gg (E \text{ produce } E)) \wedge \neg(\neg E \gg (\neg E \text{ produce } E))$$

to $E \wedge (E \gg E) \wedge \neg(\neg E \gg E)$, which is itself equivalent to E .

Alternatively, one may account for the fact that *cause* is irreflexive, but *because* is not, by claiming that *cause* requires its relata to be distinct but *because* does not; in other words, that C *cause* E is equivalent to E *because* $C \wedge (C \not\equiv E)$.⁴ However, the fact that *cause* is never reflexive, while *because* can be, appears to be part of a larger pattern, one that is not captured merely by requiring distinct relata. As is well-known, *because* has ‘noncausal’ or ‘explanatory’ uses:

- (10) a. S satisfies the axiom of extensionality *because* it is a set.
b. #The fact that S is a set *causes* it to satisfy the axiom of extensionality.
- (11) *Uttered in a situation where B is false.*
a. $A \vee B$ is true *because* A is true.
b. #The fact that A is true *causes* $A \vee B$ to be true.

S is a set is not equivalent to S satisfies the axiom of extensionality, and A is true is not equivalent to $A \vee B$ is true, but still we observe a contrast between *cause* and *because* in (10) and (11). Requiring that *because* allows the relata to be the same, while *cause* does not, therefore does not account for the differences between *cause* and *because*.

It is well-known that *because* allows for so-called ‘epistemic’ readings, as in (12a), which *cause* does not allow.⁵ Compare:

- (12) a. Dari is home *because* his lights are on.
b. The fact that Dari’s lights are on *caused* him to be home.

A relevant question is whether the differences between *cause* and *because* can be traced to an epistemic interpretation of *because*.⁶ The thought is that I *did it because* I *did it* is acceptable *because* we interpret it as I *know that* I *did it because* $(I$ *know that*) I *did it*. There is some evidence against this strategy. Compare:

⁴Lewis (1973a), for example, defines causal dependence as a relation between *distinct* events.

⁵On epistemic readings of *because*, see Kac (1972), Powell (1973), Kim (1974), and Morreall (1979).

⁶One way to implement this would be in the system of Meyer (2013), who proposes that assertions are by default interpreted with a silent knowledge operator.

- (13) a. (i) A: Why is John home?
 (ii) B: # John is home because his lights are on.
 b. (i) A: Why did you do it?
 (ii) B: I did it because I did it.

A *why* question expects an answer of the form *E because C*. A plausible account of why (13a-ii) is odd is that, instead of B responding with something of the desired form *E because C*, B responds with an unrequested form, *s knows that E because C*. If *because* had the same semantics as *cause*, but allows for reflexive readings due to a silent epistemic operator, we would expect (13b-ii) to be equally strange. However, it appears that (13b-ii) is an acceptable, if uninformative answer to A's question. This is expected if *because* does not need an epistemic interpretation to be reflexive, but can be reflexive instead by using improper chains in its notion of production.

To conclude, in this section we have seen evidence suggesting that *cause* requires the production chain to be proper, while *because* allows improper chains.

5.5 The chain's relata: situations

Our definition of chains makes essential appeal to situations, each of which represents some part of the world at a moment in time. This temporal specificity is a deliberate choice. It is motivated by overdetermination cases: cases of causation without counterfactual dependence, such as the Billy and Suzy context in (2), and cases of inevitable effects.⁷ Consider:

- (14) a. Socrates died because he drank poison.
 b. Socrates drinking poison caused him to die.

If the chain's relata were not temporally specific, and specified merely whether or not a given proposition is true at a world, then for (14) to be true we would require that the truth of *Socrates died* at a world counterfactually depends on some part of the world being the way it is. But given that Socrates' death was inevitable (given that he was born), whether or not he dies at some point is counterfactually independent of every part of the world being the way it is.

In contrast, by using temporally-specific information in the production chain, we can construct such a chain as follows: where each description is intended to pick out a particular situation: Socrates is drinking poison at time *t* ... Some poison is entering his cells at time *t'* ... his lungs are not delivering oxygen at time *t''* ... he is dead at time *t'''*. If a previous event on the chain had not occurred at the time it did, Socrates would not have died at the time he did – while truth-at-a-world fails the counterfactual dependence test here, temporally-specific situations pass it with flying colours.

⁷We discuss previous approaches to inevitable effects in section 5.6.

Compare this with the switch case. Recall (4):

- (4) a. The train reached the station because the engineer pulled the lever.
- b. The engineer pulling the lever caused the train to reach the station.

To correctly predict that (4), repeated below, are false when the tracks converge, we need to say, loosely speaking, that if the engineer had not pulled the lever, her not pulling the lever would have produced the train to reach the station. We can find such a production chain; for example, the engineer does not pull the lever at time t ... the train is at position x at time t'' ... the train is at position y at t''' ... the train reaches the station at time t'''' .

It may appear that our account of production requires that for C to produce E , it must be that if C had not occurred, E would have occurred at a different time. This, however, is not the case. Given that production involves a chain from the cause to the effect, we do not require that when the effect occurs counterfactually depends on the when the *cause* occurs; rather, when the effect occurs must counterfactually depend on a previous event on the chain, which need not be the cause itself.

It is often been pointed out that there are cases of causation where, if the cause had not occurred, the effect would still have occurred at the same time anyway (Schaffer 2000, Yablo 2004, Hall 2004).

Billy throws a Smart Rock, equipped with an onboard computer, exquisitely designed sensors, a lightning-fast propulsion system – and instructions to make sure that the bottle shatters in exactly the way it does, at exactly the time it does. In fact, the Smart Rock doesn't need to intervene, since Suzy's throw is just right. But had it been any different – indeed, had her rock's trajectory differed in the slightest, at any point – the Smart Rock would have swooped in to make sure the job was done properly.

(Hall 2004: due to Yablo, p.c.)

In this context (2) are still intuitively acceptable, while (3) are not.

- (2) a. The bottle broke because Suzy threw her rock at it.
- b. Suzy throwing her rock at the bottle caused it to break.
- (3) a. The bottle broke because Billy threw his rock at it.
- b. Billy throwing his rock at the bottle caused it to break.

We may consider a production chain of the form: Suzy throws the rock at t ... the rock is in position x at t' ... the rock is in position y at time t'' ... the rock hits the bottle at time t''' . Now consider the time one millisecond before the rock hit the bottle. I wager that there must be some time t^* , arbitrarily close to the time when the bottle broke, such that the rock was in position x^* at time t^* , and

if the rock hadn't been in position x^* at time t^* , the bottle would not have broken when it did. Now, it is highly plausible to assume that it takes some time for the Smart Rock to recognise that Suzy's rock is not on its assigned course and react accordingly. Since we require the production chain to be dense, there is a time after this window, but still on the chain, such that if it had not occurred at the time it did, the bottle would not have broken when it did.⁸

We therefore predict that Suzy throwing her rock produced it to break. And given that her throw was on-target and with the required force for the bottle to break, we also predict that Suzy throwing her rock was sufficient for it to produce the bottle to break. So the positive condition is satisfied. Finally, given that if Suzy hadn't thrown, her not throwing would not have produced the bottle to break – the negative condition is satisfied too. We therefore predict (2) to be true in the Smart Rock case.⁹ This shows that on our account, causation does not require making a difference to when the effect occurred. It is possible for C cause E to be true even though had C been false, E would have occurred at the exact same time it fact did.

5.5.1 Counterfactual dependence

Lewis (1973a) used chains of counterfactual dependence in his attempt to analyse causal dependence. Here we use it to analyse production, which is a part of the meaning of *cause* and *because* but certainly not the whole story: their meaning also involves difference-making and sufficiency. On our diagnosis, then, Lewis's proposal exhibits an accidental synecdoche: he took the part (production) for the whole (difference-making + sufficiency + production).

In the previous chapter we offered an analysis of sufficiency. In the spirit of making do with what we already have, in this chapter we will analyse production in terms of sufficiency. We will say that E counterfactually depends on C just in case $\neg C$ is sufficient for $\neg E$; in symbols: $\neg C \gg \neg E$.

Now we face the question of how exactly a chain of events is supposed to be 'held together' by a relation, such as counterfactual dependence. We turn to this question now.

5.5.2 Linking the chain

Let S be a set and \leq a binary relation over S . For any $x, y \in S$ define that $x < y$ just in case $x \leq y$ and $y \not\leq x$.

⁸One potential response to this account is that it assumes it takes some time for effects to take hold; for example, it takes some time for the Smart Rock to act in response to Suzy's rock going off course. This is tantamount to the assumption that there is no simultaneous causation. We address this concern in section 6.3.2.

⁹Similar remarks apply to the case of Merlin and Morgana from Schaffer (2000).

For any $x, y \in S$, we define that y is a *successor* of x just in case $x < y$, and that y is an *immediate* successor of x just in case $x < y$ and for no $z \in S$ is $x < z < y$. We define that \leq is *discrete* just in case for any $x, y \in S$ such that $x < y$, there is a $z \in S$ such that $x < z \leq y$ and for no $z' \in S$ is $x < z' < z$ (when \leq is linear, this boils down to requiring that every element with a successor has an immediate successor).

5.5.1. DEFINITION (Links). Let (S, \leq) be a linear order and R a binary relation over S .

1. R links (S, \leq) just in case for all $x \in S$, if x has a successor then it has a successor $y \in S$ such that for all $z \in S$, if $x < z \leq y$ then xRz .
2. R discretely links (S, \leq) just in case \leq is discrete and every element is R -related to its immediate successor: for all $x, y \in S$, if y is an immediate successor of x then xRy .

Discrete linking is more intuitive, saying that every element is related to the next element. But it only works when the chain is discrete. The general linking condition says that every element with a successor is related to a ‘buffer’ of elements in front of it. One may show that if the order is discrete the general and special definitions coincide, i.e. linking is equivalent to discrete linking (Fact 5.5.2).

5.5.2. FACT. Let (S, \leq) be a linearly ordered set and R a binary relation over S .

1. If R discretely links (S, \leq) then R links (S, \leq) .
2. If R links (S, \leq) and \leq is discrete then R discretely links (S, \leq) .

Therefore

3. If \leq is discrete, R links (S, \leq) if and only if it discretely links (S, \leq) .

PROOF. For (1), suppose (S, \leq) is discretely linked by R and pick any x with a successor. Then x has an immediate successor y and xRy . Pick any z such that $x < z \leq y$. Since y is an immediate successor of x , $z \not< y$, then as $z \leq y$, $z = y$, so xRz .

For (2), suppose \leq is connex, reflexive, and discrete and (S, \leq) is linked by R . Pick any $x, y \in S$ such that y is an immediate successor of x .

Since (S, \leq) is linked by R , x has a successor y' such that xRz for all z with $x < z \leq y'$. Since \leq is connex, either $y \leq y'$ or $y' \leq y$. If $y \leq y'$ then xRy . And if $y' \leq y$ then as $y' \not< y$, $y' = y$, then by reflexivity, $y' \leq y'$, so xRy' and so xRy .

The right-to-left direction follows from (1). \square

We propose that the chain involved in production is linked by counterfactual dependence. The counterfactual dependence is between the actuality of states,

saying that if state s hadn't been actual at t , state s' wouldn't have been actual at t' . What does it mean for a state s to be actual at a moment t ? There is a natural answer to this question: it means for s to be part of t . Thus for any state s and moment t , let us say that $actual(s_t)$ is true just in case s is part of t , and $\neg actual(s_t)$ is true just in case s is not part of t .

We analyse this notion of counterfactual dependence in terms of sufficiency, the notion we formalised in chapter 3. For any world w and situations s_t and s'_t that are part of w , $actual(s'_t)$ counterfactually depends on $actual(s_t)$ at w just in case $\neg actual(s_t)$ is sufficient for $\neg actual(s'_t)$ at w .

To apply our analysis of sufficiency from chapter 3, we need to say what states $actual(s_t)$ is about. This is also a natural answer to this question: $actual(s_t)$ is about s and no other states. Similarly, $\neg actual(s_t)$ is about s and no other states. Section 3.6.2 then tells us that $\neg actual(s_t)$ is sufficient for $\neg actual(s'_t)$ just in case when we remove s from t , every nomically possible world that contains what is left over but not s_t is one that does not contain s' at t' .

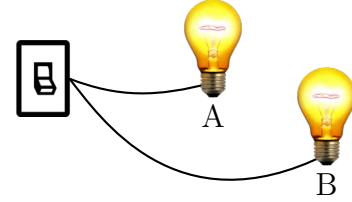
Lastly, we analyse what it means for sentence C to produce sentence E , we need to say what it means for the production chain to be “from C ” and “to E ”. Recall that our modelling framework comes with an aboutness relation between sentences and states. We propose that a chain $\{s_t\}_{t \in I}$ is *from* C just in case, where s_t is the first element of the chain and s'_t the last, s is part of a state C is about and s' is part of a state E is about. Allowing s to be part of a state C is about (rather than requiring that C is about s) allows the cause and effect to be about more of the world than strictly required by the production chain. Some evidence for this is the naturally-occurring examples from section 2.7, repeated below.

- (15) a. Computers do an awful lot of deliberation, and yet their every decision is wholly caused by the state of the universe plus the laws of nature.
- b. If anything is happening at this moment in time, it is completely dependent on, or caused by, the state of the universe, as the most complete description, at the previous moment.
- c. If you keep asking “why” questions about what happens in the universe, you ultimately reach the answer “because of the state of the universe and the laws of nature.”

Thus we have answered the four questions in section 5.3 that we required to analyse production. C produce E is true at a world w just in case there is a chain of w from C to E that is linked by counterfactual dependence between the actuality of states.

5.5.3 Evidence for this choice of aboutness: backtracking

Our choice of aboutness for $actual(s_t)$ appears to be an obvious choice. There are also data supporting it. Recall our discussion of common cause structures from section 3.6.4. Recall (16) and (17).



- (16) a. If light A were off, light B would be off.
b. If light A turned off, light B would turn off.
- (17) a. Light A turning on caused light B to turn on.
b. Light B turned on because light A turned on.

The sentences in (16) admit a true, backtracking reading while the sentences in (17) do not. The difference is that the semantics of *cause* and *because* involves production, while the semantics of conditionals does not. Light A turning on did not produce light B to turn on. To predict this, we must ensure that when we evaluate counterfactual dependence along the production chain, we keep the state of the switch fixed. This is predicted by our choice of aboutness in the analysis of production. Given that $\neg actual(s_t)$ is only about s_t , where s_t is a situation of light A being on, we do not vary the state of the switch, as desired.

5.5.4 Evidence for strong dependence: chain widening

We have proposed that production chains are linked by counterfactual dependence. We took counterfactual dependence to be strong, in the sense that for $actual(s'_t)$ to counterfactually depend on $actual(s_t)$, we require $\neg actual(s_t)$ to be sufficient for $\neg actual(s'_t)$. An alternative proposal is that the dependence to be weak, requiring $\neg actual(s_t)$ to *not* be sufficient for $actual(s'_t)$.

There is an argument that we need the strong notion of dependence. The argument is that the weak form of dependence allows for an undesirable situation I call *chain widening*. To see the issue, let us recall a classic example of a causal chain: dominoes falling in sequence. Imagine two chains of dominoes. The first domino of each chain is pushed at the same time. We want to say that for each chain, pushing the first domino of that chain produced the final domino to fall over, and did not produce the final domino of the other chain to fall over. We can show this on our analysis of production, as depicted in Figures 5.4 and 5.5, with a chain of counterfactual dependence from the state of each domino falling to the state of the next falling.

Look at the first domino of the upper chain. If it had not fallen, the second domino of the upper chain would not have fallen. But also, if the first domino had not fallen, the whole collection of dominoes would be in a different state. The chain has widened. And if the whole collection of dominoes were in a different

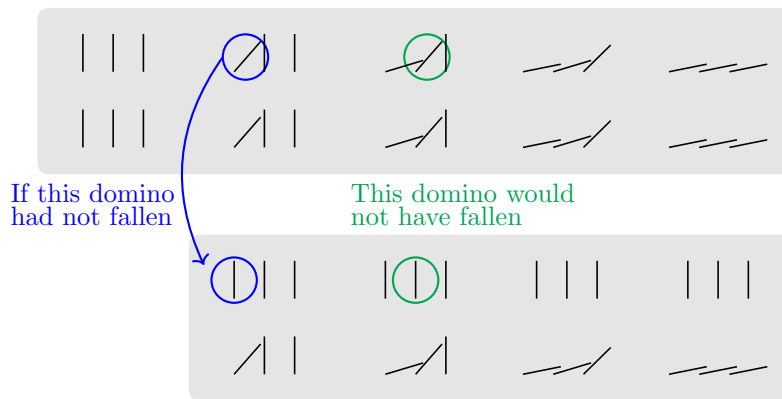


Figure 5.4

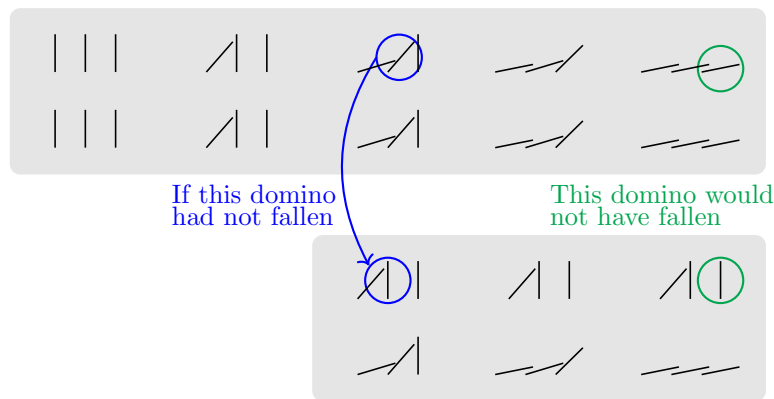


Figure 5.5: How production is supposed to work.

state, there are many possibilities to consider. In *some* of them (but not all), the lower chain is different. We see a change in the state of the lower chain, as shown in Figures 5.6 and 5.7.

If the notion of dependence involved in production were weak, the mere existence of some possibility where the other chain is different would be enough for the production chain to widen from the upper chain to the lower chain. We would then predict that the first domino of the upper chain falling produced the last domino of the lower chain to fall. This is not what we want.

More generally, we could use this chain widening strategy to create production chains between any events whatsoever, so long as the chain respects the order of time. Taking a weak notion of dependence in the analysis of production trivialises production. The strong notion of dependence, in contrast, is immune to the chain widening problem.

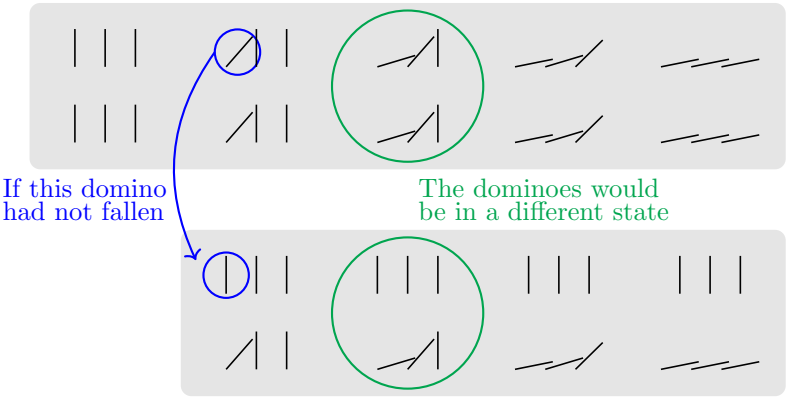


Figure 5.6: Chain widening.

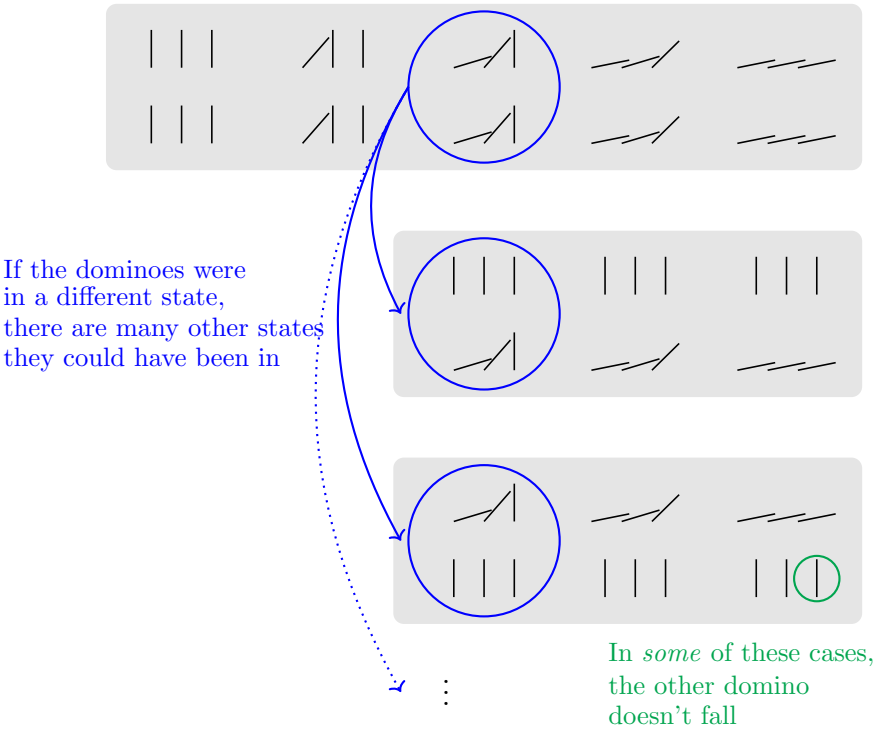


Figure 5.7: How production could work if we took dependence to be weak in the analysis of production.

5.6 Inevitable effects

In this section we compare our approach to preemption cases from chapter 2 with previous approaches. Consider:

- (18) a. Socrates drinking poison caused him to die.
 b. Socrates died because he drank poison.

Given Plato’s account in the *Phaedo*, these sentences are perfectly acceptable.

A widespread view in the literature on causation is that for a causal claim to be true there must be some possibility where the effect does not occur. In a slogan, if something has a cause then its occurrence/truth was contingent, a principle we may call *effect contingency*.¹⁰ Where the causal claim is of the form *C cause E* or *E because C* effect contingency states that, for these sentences to be true, there must be some possibility where *E* is not true. This is an informal principle, since it leaves open in what sense *E*’s falsity is taken to be impossible. Nonetheless, however we understand this notion, it is clear that the acceptability of sentences such as (18) poses a challenge for effect contingency (see e.g. Lewis 2000). Given that Socrates had to die eventually, it is hard to see how we could find a possibility where he does not die. Something must be said to correctly predict that (18) are acceptable; say, concerning what scenarios we are consider when we interpret these sentences, or how long into the future after Socrates drank poison we are allowed to look, or how we interpret the expression *Socrates died*. Call this the problem of inevitable effects.

Here are examples of the problem of inevitable effects for two recent proposals. Andreas and Günther (2021) offer an analysis of the expression *if a cause of*, proposing that “for *c* to be a cause of *e*, there must be a causal model $\langle M, V' \rangle$ that is uninformative about *c* and *e*, while intervening by $\neg c$ determines $\neg e$ to be true”, where *M* is a causal model and *V'* a partial assignment of values to variables (Andreas and Günther 2021:685, for formal details see their paper).

Now recall the switch scenario with the convergent tracks. In this scenario (19) is intuitively unacceptable.

- (19) The engineer pulling the lever is a cause of the train reaching the station.

Andreas and Günther account for this as follows.

Why is *f*’s firing [the engineer pulling the lever] no cause of *e*’s firing [the train reaching the station] in the switch scenario? Well, there is simply no causal model $\langle M, V' \rangle$ that contains no information about *e*. Even if there is no information at all as to which events occur, the

¹⁰Proposals that appeal to a possibility where the effect does not occur include Lewis (1973a), Yablo (2004), Halpern and Pearl (2005), Weslake (2015), Halpern (2016), Beckers and Vennekens (2018), Beckers (2021a), and Andreas and Günther (2020, 2021).

information about the dependences between the events is sufficient for e 's occurrence. No matter whether f or $\neg f$ is actual, e occurs according to the structural equations.

(Andreas and Günther 2021:682)

The problem is that we can say the same about Socrates' death. In any causal model representing the fact that Socrates' death was inevitable, information about the dependencies alone is sufficient for Socrates to die.¹¹ However, the following sentence is intuitively fine.

(20) Socrates drinking poison was a cause of his death.

Given Andreas and Günther's appeal to a possibility where the effect does not occur, we are left without an account of the contrast between (19) and (20).

Inevitable effects are also a problem for Beckers (2016), who also predicts (20) to be false. To apply Beckers' analysis of production we must construct a structural causal model with a variable representing whether or not Socrates dies. Now, given the laws of biology, it is natural to assume that death is inevitable. In terms of nomic possibility, we may say that in every nomically possible world where Socrates is born (as a human), Socrates eventually dies. But then everything whatsoever is sufficient for *Socrates dies* to be true.¹² In terms of Beckers' account, let E be a variable representing whether or not Socrates dies, and be C_i the penultimate event on the chain. Given that Socrates' death is inevitable, we will not be able to find a set L such that $L \setminus \{C_i\}$ is not sufficient for E – the NESS test fails. In that case nothing can produce Socrates to die, and as *cause* and *because* require that the cause produced the effect, we would wrongly predict (18) to be false.

¹¹If one responds that Socrates' dying was guaranteed not by the dependencies alone but also by some matters of fact (say, the fact that Socrates was born in the first place) note that we can say the same about the switch case: the train reaching the station is inevitable given that the train began its journey. The correspondence between the switch case and Socrates' death is still preserved. Andreas and Günther's reasoning above does not account for why (19) is unacceptable but (20) is fine.

¹²Let us show this in more detail. Beckers (2016:77) defines sufficiency in terms of structural causal models and logical entailment, whereby a set of literals L implies an atomic sentence X just in case $\bigwedge L$ logically implies the structural equation for X . In terms of structural causal models, it is plausible to formalise the claim that Socrates' death was inevitable as the claim that the structural equation for SD is a tautology, where SD is the atomic sentence representing that Socrates dies. On this formalisation of inevitability, Beckers predicts that every set whatsoever is sufficient for SD . The NESS test fails: Beckers' predicts that nothing can produce nothing can produce SD . Since causation requires production, Beckers therefore predicts that nothing is a cause of Socrates dying. One may try to formalise inevitability in a way that avoids this problem. Since we propose another analysis of sufficiency anyway (the analysis from the previous chapter), we will not pursue this project here.

5.6.1 Previous responses to the problem of inevitable effects

A popular reply to the problem of inevitable effects is that causation does require dependence not in *whether* the effect occurs, but in something about *how* the effect occurs (Paul 1998, 2000, Lewis 2000). The typical way to implement this idea is using counterfactual dependence: if the cause had not occurred, the effect would have occurred in some relevantly different way. For example:

e depends causally on *c* iff *c* occurs, *e* occurs, and if *c* had not occurred, then *e* would not have occurred at all, or would have occurred later than the time that it actually did occur.

(Paul 1998:193)

Suppose it were alleged that since we are all mortal, there is no such thing as a cause of death. Without the hanging that allegedly caused the death of Ned Kelly, for instance, he would sooner or later have died anyway. Yes. But he would have died a different death, and the event that actually was Kelly's death would never have occurred.

(Lewis 2000:185)

This is a compelling response, and I believe there is something to it. We can indeed restrict the interpretation of *to die* to something specific concerning the time or manner of the death, illustrated in (21).

- (21) Doctor: The test results just came back.
 Patient: What do they say? Am I going to die?
 Doctor: No, you're not going to die. But you do need a new prescription.

In this context "not going to die" means something like *not going to die soon*, or *from the particular illness in question*.¹³

One property of domain restrictions is that they can be cancelled, as in (22).

- (22) A: There's a 20% discount on everything.
 B: Really? On everything in the world?
 A: Well, not on *everything*. Just on everything in this shop.

Likewise, the restriction in the doctor's office example can be easily cancelled.

¹³The example also works with *will* in place of *be going to*:

- (i) Doctor: The test results just came back.
 Patient: What do they say? Will I die?
 Doctor: No, you won't die. But you do need a new prescription.

The example is arguably more natural with *be going to* than *will*, which one may account for using the differences between *will* and *be going to* observed by Copley (2001, 2008, 2009).

- (23) Doctor: No, you're not going to die. Well, you are human, so of course you are going to die, but not soon/as a result of this illness.

This is expected if (21) involves a contextual restriction; say, to a salient set of dying events.¹⁴

The question is whether such a restriction is the reason why we can say that inevitable events have causes; for example, why (18) are fine. If it is, we expect the restriction to be similarly cancellable, in which case these sentences should become unacceptable (on a counterfactual dependence approach to causation). To test this, consider:

- (24) a. Drinking poison caused Socrates to die. Well, of course he was going to die eventually, so drinking poison didn't cause him to die.
 b. Socrates died because he drank poison. Well, of course he was going to die eventually, so he didn't die because he drank poison.

These sound incoherent. One can assert in the same breath that Socrates' death was both inevitable and had a cause. This strongly suggests domain restriction is not responsible for the fact that (18) are acceptable.

Now, Paul and Lewis do not appeal to contextual domain restriction. Rather, they propose that sensitivity to the time/manner of the effect's occurrence is part of the meaning of causal terms themselves. In the following section we consider two arguments against this idea. Our first argument is that for every overdetermination case where the causal claim is true and the cause made a difference to the specific way in which the effect occurred, we can construct a switch case where the cause also made a difference to the specific way in which the effect occurred but the causal claim is false. While this objection has been around for some time (made, for example, by Rice 1999, Hall 2004, Yablo 2004), our second objection is novel, concerning the meaning of *only because*.

5.6.2 For every overdetermination case there is a switch

Recall Hall's switching context with the converging tracks (Figure 5.1). The bottom track is shorter than the top track. Let us suppose that the train reached the station before noon, but if the engineer had not pulled the lever, the train would have reached the station after noon. Pulling the lever hastened the train's arrival. Still we reject (4), repeated below.

- (4) a. The train reached the station because the engineer pulled the lever.
 b. The engineer pulling the lever caused the train to reach the station.

There is a clear difference in meaning between the sentences in (4) and:

¹⁴For an overview of contextual domain restriction, see Stanley and Szabó (2000).

- (25) a. The train reached the station before noon because the engineer pulled the lever.
- b. The engineer pulling the lever caused the train to reach the station before noon.

We might imagine that pulling the lever hastened the train's arrival by days, months or years. It does not matter: once we are given that if the engineer hadn't pulled the lever, the train would have taken the upper track and would have reached the station eventually, (4) are unacceptable.

This suggests that one can hasten the effect's occurrence without causing it. What about changing the way in which the effect occurs? Imagine a scenario with two converging tracks, where if the engineer does not pull the lever the train reaches the station from the East, but if she pulls the lever it reaches the station from the West. The engineer pulls the lever and the train reaches the station from the East. In that case, if the engineer hadn't pulled the lever, arguably it would have arrived in a different way. We might say it would have had a different arrival. In that context (4) are still unacceptable. The contrast would be puzzling if *cause* and *because* expressed counterfactual dependence in the time or manner in which the effect occurred.

We can make the same point regarding causes of death. Suppose the Athenian citizens vote to put Socrates to death, but leave it to the executioner to decide when he has to die. The executioner was planning a year-long trip to Babylon, but his boat was destroyed in a storm. Socrates died in 399 BCE, but if the executioner's boat hadn't been destroyed Socrates would have died a year later, in 398 BCE. Consider:

- (26) a. Socrates died because the executioner's boat was destroyed.
- b. The fact that the executioner's boat was destroyed caused Socrates to die.
- (27) a. Socrates died in 399 BCE because the executioner's boat was destroyed.
- b. The fact that the executioner's boat was destroyed caused Socrates to die in 399 BCE.

There is a clear contrast between (26) and (27). (26) are unacceptable even though the boat's destruction hastened Socrates' death, while given the set up above, (27) are fine.

Similarly, imagine that the executioner had only one dose of hemlock left, designated for another prisoner. The Athenians originally wished to throw Socrates off a cliff. However, the other prisoner was released, so the hemlock was given to Socrates instead. Consider:

- (28) a. Socrates died because the other prisoner was released.
- b. The other prisoner's release caused Socrates to die.

These are unacceptable, even though the other prisoner's release changed how Socrates died. Had the other prisoner had not been released, Socrates would have died in a different way; we might say, in the words of Lewis (2000), that he would have died a different death. (29), in contrast, are fine.

- (29) a. Socrates died by hemlock poisoning because the other prisoner was released.
 b. The other prisoner's release caused Socrates to die by hemlock poisoning.

One may of course reply that we have misconstrued the identity criteria of events, that once we get we have the right account of when two events are the same, we will see that in (18) if the cause had not occurred, the effect would have constituted a different event, but in (25)–(28) the effect would have constituted the same event.

A central difficulty with this response is that (18) are clearly acceptable while the identity criteria of events are much less clear. We would like an account that makes clear predictions for clear judgements. But appeals to the identity criteria of events seem to pull us toward a convoluted system of 'bells and whistles', in the words of Lewis (2000). For example, Hitchcock (2012:83) wonders:

if a meeting is originally scheduled for Monday at noon, and then re-scheduled for Tuesday at noon, is the meeting that occurs on Tuesday at noon the very same meeting that would have occurred on Monday? That is, was the meeting postponed, strictly speaking, or was the original meeting cancelled and a different meeting scheduled for Tuesday?

I find both responses plausible: we can call the new event 'the same meeting' as the original meeting or not.¹⁵ In contrast, (18), repeated below, are unquestionably true.

- (18) a. Socrates drinking poison caused him to die.
 b. Socrates died because he drank poison.

The clarity of our judgement that (18) are true, and unclarity of our judgements about when events are the same, suggests that our interpretation of *cause* and *because* is not based on the identity criteria of events, but built on firmer foundations.

In contrast to this uncertainty, Beckers' proposal gives a straightforward account of the difference between overdetermination cases and switch cases. In

¹⁵A further difficulty is that a number of theories of the identity criteria of events appeal to causal notions. For example, Davidson (1969) proposes to individuate events by their causes and effects. We might like to appeal to such accounts to help with the identity criteria of events, but cannot do so here on pain of circularity.

overdetermination cases the absence of the cause would not have produced the effect, while in switch cases it would have. For example, if Socrates had not drunk poison, him not drinking poison would not have produced his death; rather, something else would have (drinking too much wine, fighting the Spartans, ...).

Our second argument against the idea that causal claims express some kind of counterfactual dependence in how the effect occurs concerns the meaning of *only because*.

5.6.3 *Only + because = counterfactual dependence*

As we have seen, a common response to overdetermination cases on behalf counterfactual dependence approaches to causation is to say that causation involves counterfactual dependence in *how* the effect occurs, rather than *whether or not* it occurs. The strategy, generally speaking, is to amend the notion of counterfactual dependence to find counterfactual dependence where at first sight there is none. For example, given that Socrates' death was inevitable, the fact that he died does not counterfactually depend on anything, so Paul, Lewis, and others reply that the counterfactual dependence is of a different sort.

This strategy backfires in an interesting way when we look at *only because*. Recall that the following sentences are fine:

- (30) a. The bottle broke because Suzy threw her rock at it.
 b. Socrates died because he drank poison.

Now consider:

- (31) a. The bottle broke *only because* Suzy threw her rock at it.
 b. Socrates died *only because* he drank poison.

These sentences have (at least) two readings. On one they imply counterfactual dependence:

- (32) a. If Suzy hadn't thrown her rock, the bottle would not have broken.
 b. If Socrates hadn't drunk poison, he wouldn't have died.

On this reading (31) are unacceptable. A second reading is that the effect did not happen because of anything else, which we may make salient as follows.

- (33) a. The bottle didn't break because Billy threw his rock at it, and it didn't break because Charlie hit it with a hammer. The bottle broke *only because* Suzy threw her rock at it.
 b. Socrates didn't die because he drank too much wine, and he didn't die because the Spartans attacked him. Socrates died *only because* he drank poison.

On this reading (31) are acceptable.

Here we are interested in the counterfactual dependence reading. It is clear why (31) are unacceptable on this reading: in the actual circumstances, the effect did not counterfactually depend on the cause. If Suzy hadn't thrown her rock the bottle would have broken anyway. If Socrates hadn't drunk poison he would have died anyway. To test this, we can remove Billy from the scenario, so that the bottle breaking counterfactually depended on Suzy's throw. In that case suddenly (31a) becomes fine. Likewise, if we imagine that Socrates would have lived forever had he not drunk the poison, (31b) is perfectly acceptable.

The problem for counterfactual dependence approaches to causation, such as Paul (1998) and Lewis (2000), is that, to account for (31)'s unacceptability, we need to appeal to counterfactual dependence – the original notion concerning whether or not the effect occurred. But counterfactual dependence approaches to causation have analysed this notion away in favour of amended version concerning the specific way in which the effect occurs.

Let's see what happens when we explicitly state that the effects are temporally specific, as in (34).

- (34) a. The bottle broke before time t only because Suzy threw a rock at it.
 b. Socrates died before time t only because he drank poison.

There are times for which these sentences are perfectly acceptable. If *cause* and *because* expressed counterfactual dependence in when the effect occurred, we would expect no difference between (31) and (34), for suitable choices of t . This is not what we observe; rather, (31) are much worse than (34).

We find the same pattern if we replace temporal specificity with specificity in the way in which the effect occurred, as in (35).

- (35) a. The bottle broke in way x only because Suzy threw a rock at it.
 b. Socrates died in way x only because he drank poison.

Again, (31) are much worse than (35), a fact which is unexpected if the counterfactual dependence involved in *cause* and *because* involved counterfactual dependence in the way the effect occurred.

One reply on behalf of counterfactual dependence approaches would be to say that *only* modulates between different notions of counterfactual dependence; specifically, saying that *only* turns *how*-counterfactual dependence into *whether*-counterfactual dependence. However, there is nothing in the meaning of *only* to suggest this ability. Considerations of compositionality favour an account of the meaning of *only because* in terms of the meaning of *only* and the meaning of *because*.

5.6.4 *Only because* on the present account

The present account of the meaning of *because* can account for unacceptability of (31). Moreover, it predicts the counterfactual dependence reading of *only because* in a compositional way: the reading falls out of our semantics of *cause* and *because* together with a standard semantics for *only*.

First we need a semantics of *only*. Recall our discussion from section 2.7.2. It is well-known that *only* is interpreted with respect to a set of alternatives (Rooth 1985), as in the following standard entry, which we adopt here.¹⁶

- (36) **Meaning of *only*.** For any sentence S and set of sentences Alt , $only_{Alt} S$ asserts that for every $A \in Alt$, if S does not entail A then A is false.

To illustrate with the classic example, compare:

- (37) a. I only introduced BILL to Sue.
b. I only introduced Bill to SUE.

In (37a), *only* negates alternatives of the form *I introduced x to Sue*, saying I didn't introduce anyone but Bill to Sue, while in (37b) it negates alternatives of the form *I introduced Bill to x* , saying that I didn't introduce Bill to anyone but Sue.

Similarly, we can trace the difference between the two readings of (31) to different alternatives associating with *only*. One available set of alternatives results from focus on the whole *because*-clause, which could be triggered, say, by the following questions under discussion.

- (38) a. Why did the bottle break?
b. Why did Socrates die?

We propose that when *only because* receives a counterfactual dependence reading, the alternatives are those that result from changing the *because*-clause:

- (39) a. {The bottle broke because x : x is a sentence}
b. {Socrates died because x : x is a sentence}

There is independent evidence for this choice of alternatives in other environments. For example, von Stechow (1997:28, taking up an idea by Roger Schwarzschild) and Vostrikova (2018) suggest that *only if* has as alternatives the set of all *if*-clauses.

A second set of alternatives is available, which can be triggered by narrow focus (say, on *Suzy* or *poison*), or questions such as:

- (40) a. The bottle broke because who threw a rock at it?
b. Socrates died because he drank what?

¹⁶Our predictions in this section also follow on Fox's (2007) entry for *only*.

These questions are associated with alternative sets such as:

- (41) a. {The bottle broke because x threw a rock at it : x is a person}
 b. {Socrates died because he drank x : x is an entity}

With these alternatives we derive the readings in (33).

Before turning to Suzy and Socrates, let us see how our proposal handles the following example. Suppose Reyna was born in Copenhagen and received a Danish passport and consider:

- (42) Reyna received a Danish passport because she was born in Copenhagen.

This sentence is acceptable.¹⁷ Now replace *because* with *only because*:

- (43) Reyna only received a Danish passport because she was born in Copenhagen.

This sentence has a counterfactual dependence reading, especially salient with narrow focus on *Copenhagen*, according to which Reyna would not have received a Danish passport had she been born outside Copenhagen. Since there is nothing special about Copenhagen compared to anywhere else in Denmark when it comes to receiving Danish passports, on this reading (43) is intuitively unacceptable.

Our proposed semantics for *because*, combined with *only*, correctly predicts this. For among the alternatives to ... *because Reyna was born in Copenhagen* we have ... *because Reyna was born in Denmark*. Now, (44a) does not entail (44b). For example, in a world where only those born in Copenhagen receive Danish passports, (44a) is true but (44b) false.

- (44) a. Reyna received a Danish passport because she was born in Copenhagen.
 b. Reyna received a Danish passport because she was born in Denmark.

(43) therefore asserts that (44b) is false. On our semantics of *because* this amounts to the following, where D says that Reyna was born in Denmark, E that she received a Danish passport.

$$\begin{aligned} & \neg(E \text{ because } D) \\ \Leftrightarrow & \neg\left(D \wedge (D \gg (D \text{ produce } E)) \wedge \neg(\neg D \gg (\neg D \text{ produce } E))\right) \\ \Leftrightarrow & \neg D \vee \neg(D \gg (D \text{ produce } E)) \vee (\neg D \gg (\neg D \text{ produce } E)) \end{aligned}$$

This disjunction turns out to be false. The first disjunct is false since Reyna was in fact born in Denmark. The second disjunct is false because Reyna being born

¹⁷It is based on the following naturally occurring example (discussed by McHugh 2020).

- (i) Reyna was born at Royal Bolton Hospital but received a Danish passport because her mother was born in Copenhagen. [\[Source\]](#)

in Denmark is sufficient for that to produce her to receive a Danish passport. And the third disjunct is false since, if Reyna hadn't been born in Denmark, not being born in Denmark would not have produced Reyna to receive a Danish passport (either she wouldn't have received one at all, or it would have been produced by something else; say, one of her parents being born in Denmark). Since the whole disjunction is false, and (43) implies it, we therefore correctly predict (43)'s unacceptability.

To test this account, imagine that the effect did in fact counterfactually depend on the cause. That is, imagine the passport rules changed so that if Reyna hadn't been born in Copenhagen she wouldn't have received a Danish passport. Now the second disjunct $\neg(D \gg (D \text{ produce } E))$ is true: Reyna being born in Denmark was not sufficient for that to produce her to get a Danish passport. For she could have been born in Denmark but outside Copenhagen, in which case she wouldn't have received a Danish passport, so nothing would have produced her to receive a Danish passport (production is factive, $C \text{ produce } E$ entails $C \wedge E$, so $\neg E$ entails $\neg(C \text{ produce } E)$). In this case, then, we predict that the whole disjunction is true and therefore (43) to be true. This is exactly what we observe. Assuming Reyna wouldn't have received a Danish passport had she been born outside Copenhagen, (43) is perfect.

Here is one way to think of what is going on with *only because*. The places in Denmark outside Copenhagen are 'backup causes': alternative causes that could have produced Reyna to receive a Danish passport, had she been born outside of Copenhagen. Put in these terms, given the alternatives in (39), and where \Rightarrow denotes entailment, we have just shown that for any sentence B ,

$$\begin{array}{llll} \text{if} & E \text{ because } C & \not\Rightarrow & E \text{ because } (C \vee B), \\ \text{then} & E \text{ only because } C & \Rightarrow & \neg(E \text{ because } (C \vee B)). \end{array}$$

In other words, *only because* asserts that the effect had no backup causes. And where there are no backup causes there is counterfactual dependence. This derives the counterfactual dependence reading of *only because*.

With a different set of alternatives (43) can be fine. For instance, suppose the set of alternatives consists of non-overlapping places: Aarhus, Berlin, Copenhagen With these alternatives (43) asserts that Reyna did not receive a Danish passport because she was born in Aarhus, that she didn't receive one because she was born in Berlin, and so on, which are all true since Reyna wasn't born in any of these places. The alternative *Reyna did not receive a Danish passport because she was born in Copenhagen* is not negated since it is entailed by the prejacent (indeed, it *is* the prejacent).

The same reasoning applies to *only because* in overdetermination cases, such as (31a), repeated below.

(31a) The bottle broke only because Suzy threw her rock at it.

When the alternatives are all other *because* clauses, as in (39a), then among the alternatives to (31a) we have (45).

(45) The bottle broke because Suzy or Billy threw a rock at it.

The prejacent, *The bottle broke because Suzy threw her rock at it*, does not entail (45), so *only* negates (45). On our semantics of *because* this amounts to:

$$\begin{aligned} & \neg(\textit{Suzy or Billy throw}) \\ \vee & \neg((\textit{Suzy or Billy throw}) \gg ((\textit{Suzy or Billy throw}) \textit{ produce bottle break})) \\ \vee & (\neg(\textit{Suzy or Billy throw}) \gg (\neg(\textit{Suzy or Billy throw}) \textit{ produce bottle break})) \end{aligned}$$

As in the Copenhagen case, each disjunct is false. Suzy or Billy did throw; the fact that either threw is sufficient for that to produce the bottle to break; and if neither had thrown, then the fact that neither threw would not have produced the bottle to break (the bottle would not even have broken, so given that production is factive, nothing would have produced it to break). Since (31a) asserts this disjunction, which is false, we correctly predict (31a) to be unacceptable.

If we instead imagine that the bottle breaking counterfactually depended on Suzy's throw (say, Billy's throw would have missed), the second disjunct becomes true: Suzy or Billy throwing is not sufficient for that to produce the bottle to break. For one way for Suzy or Billy to throw is for Billy but not Suzy to throw. Given our assumption that the bottle breaking counterfactually depended on Suzy's throw, if only Billy had thrown the bottle would not have broken; a fortiori, nothing would have produced it to break (since production is factive). So when the bottle breaking counterfactually depends on Suzy's throw, we correctly predict (31a) to be acceptable.

The second reading of (31a) discussed in (33), where (31a) is acceptable, results from the unavailability of the alternative *The bottle broke because Suzy or Billy threw a rock at it*. Now, the following alternative may still be available:

(46) The bottle broke because Billy threw a rock at it.

When it is, we predict that (31a) implies that (46) is false, which is perfectly fine since (46) actually is false – a fact we predict since Billy did not produce the bottle to break.

Just as we considered backup causes in the previous two scenarios (Reyna being born in Denmark, Suzy or Billy throwing), we may consider the set of backup causes of Socrates' death: the set of conditions such that, had Socrates not drunk poison, one of them would have produced his death instead.¹⁸ The backups may include old age, drinking too much wine at the Symposium, fighting

¹⁸Note that if Socrates had not drunk poison, it may not be determined at that moment which one of the backups would have produced his death. The set of backups contains all these potential producers of his death, had he not drunk poison.

the Spartans in the Peloponnese, and so on. Let *backup* express that one of these backup producers of Socrates' death occurred; for example, *backup* = (*Socrates drinks ten amphorae of wine* \vee *Socrates falls off a cliff* $\vee \dots$).

Now, (47a) does not entail (47b).¹⁹ For there is a logically possible world with different laws where, say, only poison can kill Socrates: in such a world, had he not drunk poison, he would have lived for ever. In such a world (47a) can be true while (47b) is false.

- (47) a. Socrates died because he drank poison.
 b. Socrates died because he drank poison or a backup occurred.

(31b) therefore asserts that (47b) is false (when (47b) is an available alternative).

$$S \text{ die only because poison} \quad \Rightarrow \quad \neg(S \text{ die because } (poison \vee backup))$$

On our semantics of *because*, $\neg(S \text{ die because } (poison \vee backup))$ is equivalent to

$$\begin{aligned} & \neg(poison \vee backup) \\ & \vee \neg((poison \vee backup) \gg (poison \vee backup) \text{ produce } S \text{ die})) \\ & \vee (\neg(poison \vee backup) \gg (\neg(poison \vee backup) \text{ produce } S \text{ die})) \end{aligned}$$

Actually, each disjunct is false. The first disjunct is false since Socrates did in fact drink poison. The second is false since Socrates drinking poison or a backup occurring is sufficient for one of these to produce him to die. The third disjunct is false since, if neither the original cause nor a backup had occurred, their absence would not have produced Socrates' death.²⁰

As in Denmark and Suzy cases above, we can test this explanation by imagining that Socrates' death did counterfactually depend on drinking poison. Imagine that hemlock was Socrates' kryptonite, so to speak. In that case the second disjunct is true: Socrates drinking poison or a backup occurring is not sufficient for that to produce him to die. For one way for that disjunction to occur is for a backup to occur; say, Socrates falls off a cliff.²¹ Given that Socrates would not have died had he not drunk poison, in that case he does not die (so nothing

¹⁹At least, the entailment does not hold when we read (47b) as *D because (C \vee B)*. Another scope disambiguation of (47b) has disjunction above *because*: *(D because C) \vee (D because B)*. On this reading (47a) does entail (47b). Here we are concerned with the first reading, where *because* scopes above disjunction.

²⁰If neither the original cause nor a backup had produced Socrates' death, it is hard to imagine what would have produced his death instead, given that the backups are supposed to cover all other potential producers of his death. Whatever we say in response to this, it is clear that Socrates not drinking poison, nor the backups occurring, would not have been sufficient to produce his death.

²¹In this case the label 'backups' is a misnomer. For they are no longer in fact backup producers of his death, but were only backups under our previous assumption that Socrates' death did not counterfactually depend on him drinking poison.

produces him to die). So assuming his death counterfactually depended on him drinking poison we predict (31b) are fine, which is the intuitively correct result.

Our account of above of (31b)'s unacceptability (given the alternatives in (39b)) made essential appeal to the fact that *Socrates died* in (31b) does not encode the time or manner in which he dies. If we replace *Socrates died* with *Socrates died at time t /in way x* , (31b) instead asserts:

$$\begin{aligned} & \neg(\textit{poison} \vee \textit{backup}) \\ \vee \neg & ((\textit{poison} \vee \textit{backup}) \gg (\textit{poison} \vee \textit{backup}) \textit{ produce Socrates die at time } t/\textit{in way } x)) \\ \vee & (\neg(\textit{poison} \vee \textit{backup}) \gg (\neg(\textit{poison} \vee \textit{backup}) \textit{ produce Socrates die at time } t/\textit{in way } x)) \end{aligned}$$

The second disjunct is true: Socrates drinking poison or a backup occurring is not sufficient to produce him to die at the time or in the manner in which he in fact did. For one way for him to drink poison or for a backup to occur is for a backup to occur and Socrates to not drink poison. In that case he would have at a different time/in a different way. In the words of Lewis (2000), he would have died a different death. Now, if the second disjunct were true then the whole disjunction would be true and we would lose our account of (31b)'s unacceptability.

In a nutshell, the challenge for counterfactual dependence approaches is this. Recall that intuitively (30b) is acceptable but (31b) is not (given the alternatives in (39b)).

(30b) Socrates died because he drank poison.

(31b) Socrates died only because he drank poison.

Counterfactual dependence approaches have to tell some story about why (30b) is fine in the absence of any counterfactual dependence from the cause to the effect. For if Socrates hadn't drunk poison, he still would have died. Now, as we have seen, (31b) is unacceptable precisely because of this lack of counterfactual dependence. The response that there is counterfactual dependence after all (say, concerning the time or manner of Socrates' death) accounts for (30b)'s acceptability at the cost of wrongly predicting (31b) to also be acceptable.

In contrast, the present approach simultaneously accounts for acceptability of (30b) and unacceptability of (31b). This is because the present approach does not need to appeal to any possibility where the effect does not occur: (30b) instead requires that Socrates not drinking poison is not sufficient for that to produce him to die. This can hold even when the effect is inevitable. (In that case something else something other than Socrates not drinking poison would have produced him to die.) Given that we did not need to analyse the inevitability away to account for (30b)'s acceptability, we are free to use that inevitability to account for (31b)'s unacceptability.

5.6.5 *Only because* an evidence for sufficiency

We have already seen evidence from sufficiency violations – such as the robot context – that *because* expresses sufficiency. These data from *only because* provide further evidence that *because* expresses sufficiency. For example, Beckers’ semantics of *is a cause of* lacks a sufficiency requirement. Let’s see what we predict for *only because* when we take Beckers’ proposal as a semantics of *because* (something Beckers does not propose), whereby *E because C* means that *C* is true and produced *E*, and $\neg C$ is not sufficient for it to produce *E*.

Take a case where the effect counterfactually depends on the cause: Suzy threw a rock at the bottle, breaking it, and Billy was not going to throw. As discussed, in that case (31a) is intuitively fine. Recall that it has following implication, assuming the alternatives in (39a):

- (48) a. The bottle only broke because Suzy threw a rock at it. = (31a)
 b. $\Rightarrow \neg(\text{The bottle broke because Suzy or Billy threw a rock at it}).$

On Beckers’ semantics – applied to *because* – (48b) amounts to:

$$\neg(S \vee B) \vee \neg((S \vee B) \text{ produce } E) \vee (\neg(S \vee B) \gg (\neg(S \vee B) \text{ produce } E))$$

The first and third disjuncts are false for the same reasons as before. Crucially, however, the second disjunct also is false. Suzy or Billy throwing did produce the bottle to break. But then (31a) implies something false, so we predict it to be unacceptable. In fact, when the bottle breaking counterfactually depends on Suzy throwing, (31a) is fine.

One might reply that in this case we should rethink the claim that Suzy or Billy throwing produced the bottle to break. Could we redefine production to have this come out false? We could, but this runs into another problem. Take the original Billy and Suzy case, where the bottle breaking did not counterfactually depend on Suzy’s throw. In that case (31a) is intuitively unacceptable. But now the second disjunct, $\neg((S \vee B) \text{ produce } E)$, is true, so we would lose our account of (31a)’s unacceptability.

Either way, then, Beckers’ semantics applied to *because* makes incorrect predictions for *only because*.²² As we have seen, when we replace production with sufficiency for production – that is, replace *C produce E* with $C \gg (C \text{ produce } E)$ in the semantics of *because* – we make correct predictions for *only because*; specifically, we correctly predict *only because* to be acceptable just in case the effect counterfactually depends on the cause. This shows the important role sufficiency plays in our account of *because*. It also shows that sufficiency is part of the literal meaning of *because* – rather than, say, an implicature – given the standard assumption that *only* negates the literal meaning of its alternatives. For example:

²²This of course is not a problem for Beckers’ own proposal, which is intended as an analysis of the expression *is a cause of*.

- (49) a. Only Sami talked to some of the students.
 b. Only Charity talked to Simon or Maher.

(49a) implies that the others did not talk to any of the students, not that they talked to none or all of the students (at least, it does not imply this without special ‘denial’ intonation). Similarly, (49b) implies that the others did not talk to Simon and that they did not talk to Maher: *only* negates the literal, inclusive meaning of *or*.²³

To conclude, evidence from *only because* suggests that sufficiency is part of the semantics of *because*. The fact that sufficiency helps give the right predictions for a complex environment – under *only* – is an unexpected and welcome result.

²³A straightforward explanation of this is that calculating an implicature under *only* would lead to a weaker meaning, making the utterance less informative (see e.g. Chierchia 2013:106, Fox and Spector 2018).

6.1 Where we are

The previous chapters proposed a semantics of *cause* and *because*. In this chapter we turn our attention to the models in which we expressed this semantics. The goal of this chapter is to explain why I chose the modelling framework I did – to hopefully give the reader a sense of its generality, simplicity, and advantages over alternative frameworks.

Before doing so, let us step back and situate ourselves within the broader narrative of the dissertation. Recall the two questions with which we began.

The modelling question. What kind of information do we use when we judge that a causal claim is true?

The meaning question. What are the truth conditions of causal claims?

Since causal claims come in many forms – too many to analyse in one dissertation – we narrowed our inquiry from causal claims in general to sentences containing the words *cause* and *because* in particular. Chapter 2 proposed an answer to the meaning question in terms of two relations, sufficiency and production. Chapter 3 gave an analysis of sufficiency, and chapter 5 an account of production in terms of sufficiency. On our account, then, the meanings of *cause* and *because*, rich as they are, boil down to a single notion: sufficiency.

So the previous three chapters gave an answer to the meaning question. In answering the meaning question, we have also answered the modelling question, since truth conditions are always interpreted relative to a model. And as the meanings of *cause* and *because* come down to sufficiency, our answer to the modelling question is that the kind of information we use to evaluate a sentence containing *cause* or *because* is the same information we use to determine sufficiency. In chapter 3 we saw that, with respect to a language (that is, a set of sentences) \mathcal{L} , a model of sufficiency has the following components:

$$(S, \leq, \mathcal{A}, P, |\cdot|)$$

where

(S, \leq) is a state space,

\mathcal{A} is an aboutness relation between sentences and states,

P is the set of nomically possible worlds, and

$|\cdot|$ assigns to each sentence the set of worlds where it is true.

Figure 6.1 illustrates how we have ultimately analysed the meaning of *cause* and *because* in terms of this model.

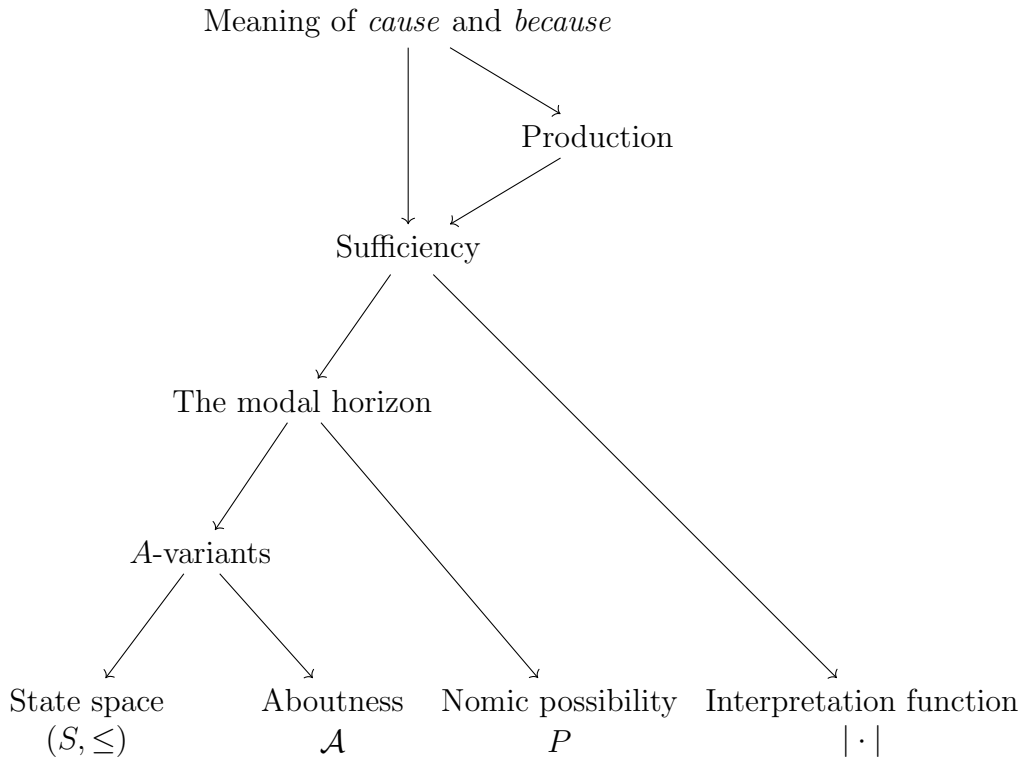


Figure 6.1

6.1.1 A formal model construction

In this model we assume states as primitive and construct worlds from them. Figure 6.2 illustrates the construction.

A *state space* (S, \leq) is a partially ordered set, the elements of which we call *states*, with \leq representing parthood. We assume that every state is part of a maximal state with respect to parthood: $\forall s \in S \exists t \in S : s \leq t \wedge \forall u \in S (t \leq u \Rightarrow t = u)$. We take states to represent snapshots, representing how some things stand at a moment in time. We can construct worlds from states as follows.

- A *situation* is a particular instance of a state.

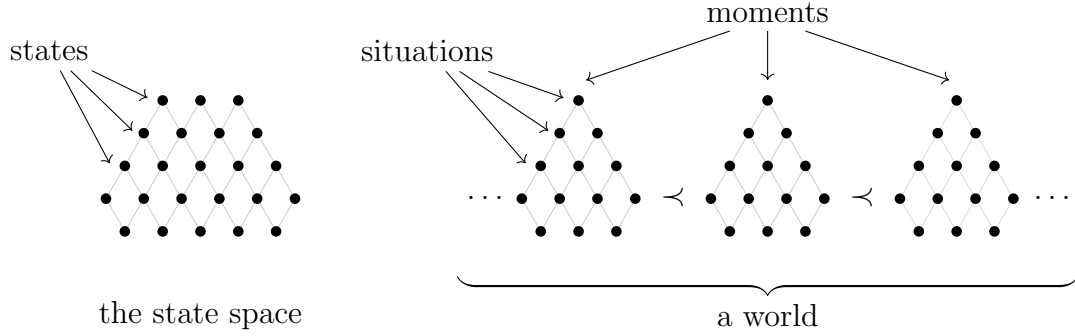
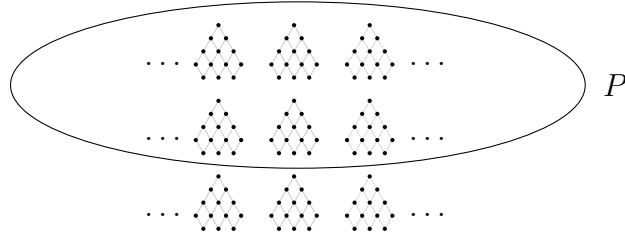


Figure 6.2: The relationship between states, situations, moments and worlds.

Figure 6.3: P is a set of worlds.

- A *moment* is a situation that is not part of any other situation.¹
- A *world* is a linear order of moments.

Formally, the set of situations, moments and worlds are defined, respectively, as follows.

$$\begin{aligned}
 Sit &:= S \times I, \text{ where } I \text{ is an arbitrary label set,} \\
 M &:= \{t_i \in Sit : t \leq u \text{ implies } t = u \text{ for all } u \in S\}, \\
 W &:= \{(M', \preceq) : M' \subseteq M, \preceq \text{ is a linear order}\}.
 \end{aligned}$$

We take states to be multiply realisable while situations are particulars. We need situations since the same state may appear multiple times in a world; for example, two objects may be in the same state, or the world may exhibit recurrence (as shown by Poincaré's 1890 celebrated recurrence theorem for dynamical systems).

6.1.2 On duration in causal models

Our representation of time specifies in what order situations occur, but not their duration. This differs from some other modelling frameworks that represent time, such as dynamical systems, which represent time with real numbers. It is always





¹We define that situations inherit parthood relations from their states; that is, situation s_i is part of situation t_j just in case state s is part of state t .

possible to impose a measure on our notion of worlds if desired. However, the fact that we can analyse sufficiency in terms of a model that does not specify duration shows that to give the meaning of causal claims, representing duration is unnecessary. Recall that, in our analysis of sufficiency from chapter 3, we needed the order of time to distinguish between the past and future of the intervention time. We did this since sufficiency treats the past and future differently, fixing the past of intervention time but allowing the future to vary. We need an order, but not duration, to distinguish the past from the future, so we need an order, but not duration, to analyse sufficiency, and therefore the meaning of *cause* and *because*.

Leaving out duration brings our model closer to our actual experience of the world. We experience the passage of time directly, but not its numerical measurement. (If we see two events that happen at different times, it is very easy to tell which one happened before the other, but not the duration between them.) Our experience of duration does not come from some absolute, Newtonian time, but from experiencing objects in the world: seeing the sun rise and set, the seasons change, the face of a clock, and so on. We can represent this on the present approach by having time-keeping devices – such as the sun and clocks – in our state space.

6.1.3 Illustrating the model construction

Here is a simple example illustrating the modelling framework, repeated from chapter 3. Consider a switch connected to a light.² When the switch is down, the light turns on, and when the switch is up it turns off. For simplicity we will ignore all other components, such as the wire and electricity in the building. We will assume there is electricity, that the bulb is in working condition and so on. We will also assume that the time is discrete and that changes take place after one step in time. Abstracting away everything except the switch and light, our state space is given in Figure 6.4.

The moments are the maximal elements:  ,  ,  and  . A world is a linear order of moments. For example, Figure 6.5 illustrates a segment from a nomically possible world.

This sequence represents, say, someone walking into a dark room and flicking the switch. The light then turns on. On their way out, they flick the switch again and the light turns off.

Figure 6.6 gives an example of a nomically impossible world, in which the light flickers on and off at random.

²McHugh (2018) presents an earlier – and less developed – formalisation of nomic possibility for this example.

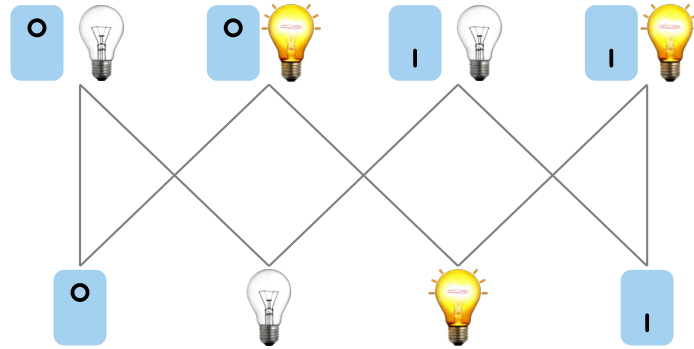


Figure 6.4: A state space of the switch and light.



Figure 6.5: A nomically possible world.

We can represent the full set of nomically possible worlds as the possible paths in Figure 6.7.

The loops show that a world where the switch is always up and the light is always off is nomically possible, as well as a world where the switch is always down and the light always on (since we are abstracting away from everything except the switch and light, we are ignoring how the system was initially set up, what happens if there is a power outage or the bulb breaks, and so on). The diagonal lines represent someone who flicks the switch at the exact same time the light changes.

We take a world to be an *infinite* directed path through the figure above to represent the fact that whenever the switch is flicked, the light eventually changes. If we allowed worlds to be finite directed paths, we could have a world where the switch is flicked and then abruptly ends. This would falsify the claim that every change in the switch leads to a change in the light. If one wished to drop this claim and allow the world to abruptly end, one can simply drop the restriction to infinite paths.

6.2 Partial models and model abstraction

Research on causal models in computer science, statistics and artificial intelligence tends to take a local view, designing bespoke models from scratch for the purpose at hand (e.g. Spirtes, Glymour, and Scheines 1993, Pearl 2000). On the other hand, research in linguistics, logic and philosophy tends to take a global view (as in possible worlds semantics), beginning with a logical space – a maximally



Figure 6.6: A nomic impossibility world.

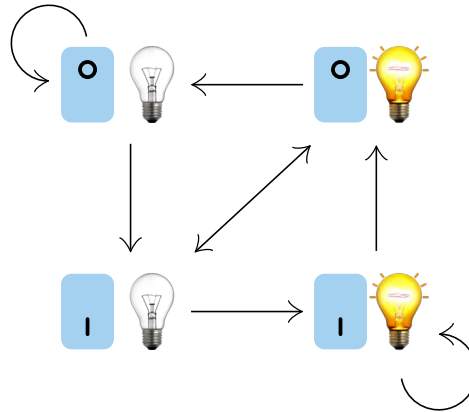


Figure 6.7: A world is nomically possible just in case it is an infinite directed path through the figure above.

inclusive space of possibilities – and understanding meanings in terms of this logical space.³

Each perspective has its pros and cons. The global view is not – and was never intended to be – psychologically plausible.⁴ There is just no way humans carry an exhaustive set of possible worlds in their heads. The local view promises to be more psychologically plausible: it is more concrete, since we can construct and inspect the models directly, which we cannot do for maximal objects in all their complete detail.

On the other hand, on the local view it is not immediately obvious how the various models relate to each other. Given many photos of a landscape, how do we patch them together to form a panorama? Given two models viewed from a local perspective, how do we know whether and to what extent they agree – whether they are contradictory, for example, or whether one is a special case of the other?⁵ The global view answers these questions easily, since the single all-encompassing model serves as

³For classic expositions of truth-conditional semantics see Davidson (1967b) and Lewis (1970b).

⁴For discussions of the psychological implausibility of semantics, see Fodor, Fodor, and Garrett (1975), Partee (1979), and Johnson-Laird (1982).

⁵For work addressing these questions with the framework of structural causal models, see e.g. Chalupka, Eberhardt, and Perona (2016, 2017), Rubenstein et al. (2017), Beckers and Halpern (2019), Beckers, Eberhardt, and Halpern (2020), Beckers (2021b), Geiger et al. (2021), and Otsuka and Saigo (2022).

a common forum for comparison (being contradictory means having no element in common, entailment means set inclusion, and so on).

One way to take the best of both perspectives is to provide a relation between models, stating when one is a high-level representation of another. This provides a general way to zoom in and out, allowing us to choose the level of detail as we please. The goal of this section is to define abstraction relation between the models we use in the present framework, which specify nomic possibilities and impossibilities.

To illustrate, recall our example with the switch and light. When discussing that example above, we made the simplifying assumption that the switch and light were the only things in the world: that a world was just a sequence of the states of the switch and light. Really, our model described the behaviour of a small part of the world.

In our model construction we defined a world to be a linear order of moments, i.e. maximal situations with respect to parthood. If we relax the maximality requirement, we get a partial version of worlds, which we call *paths*.

6.2.1. DEFINITION (Path). A path is a linearly ordered set of situations.

Every world is a path, but some paths are not worlds. For example, given a state space whose situations include more than just the switch and light, Figure 6.5 depicts a path but not a world.

A path describes how some things change through time. Sometimes we wish to ignore some of those changes. To illustrate, suppose that while the light was on, there was a noise. Figure 6.8 represents the sequence of states of the switch and the light above, expanded to specify whether or not there was noise.



Figure 6.8

We may wish to ignore the noise, to concentrate solely on the relationship between the switch and light. The changes in noise do not matter. To abstract away from the noise, we introduce an equivalence relation over situations, where states are equivalent just in case they are adjacent and differ only in whether or not there is noise. Formally, adjacency means we require that the equivalence relation be *temporally convex*: if s is equivalent to t then any situations between s and t in time are also equivalent to s and t . Convex equivalence relations partition paths into non-empty intervals (where an interval is a convex set of situations: for a set of situations I is an *interval* just in case for all situations s, t, u , if s and u are in I and $s \preceq t \preceq u$ then t is also in I), as shown in Figure 6.9.



Figure 6.9: Partitioning a path into intervals.

Where $p = (A, \preceq)$ is a path, its partition by an equivalence relation \sim , denoted $p/\sim = (S/\sim, \preceq/\sim)$ is defined as usual.⁶

We would now like to define an abstraction relation between paths, specifying when one path is an abstraction of another; in other words, when one path can be seen as a high-level description of a low-level path. Formally, we can achieve this by taking the equivalent states to represent a single state; that is, taking their equivalence classes of situations.

How do we decide whether a high level path h is an abstraction of a low-level path l ?

6.2.2. DEFINITION (Path abstraction). Given a state space (S, \leq) , path $h = (S_h, \preceq_h)$ is an *abstraction* of path $l = (S_l, \preceq_l)$ with respect to (S, \leq) just in case there are $\sim \subseteq S_l \times S_l$ and $f : l/\sim \rightarrow h$ such that

- \sim is a temporally convex equivalence relation
- f is an order isomorphism
- $f([s]) \leq s$ for all $s \in S_l$.

Let us say that l is a *refinement* of h just in case h is an abstraction of l .

Figure 6.10 illustrates that our original path without the noise is an abstraction of the path with the noise.

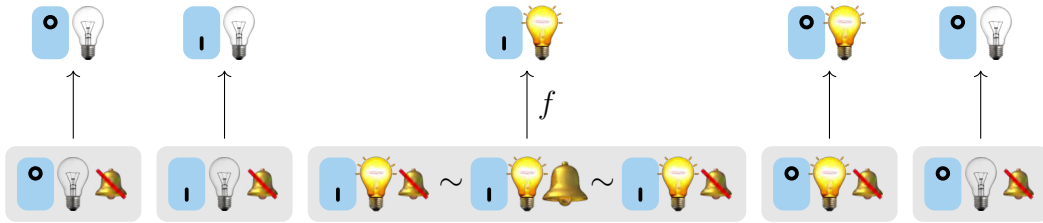


Figure 6.10: Abstracting away the noise.

Informally, a high-level path is an abstraction of a low-level path just in case there is a way of partitioning the low-level path into intervals and a way of matching each interval with a high-level state. The order isomorphism condition

⁶That is, we let $[s] = \{t \in S : s \sim t\}$ denote the equivalence class of s , i.e. the set of states equivalent to s . We then define $A/\sim = \{[s] : s \in A\}$ to be the set of equivalence classes of situations in A , and temporal succession is given by $[s] \preceq/\sim [t]$ just in case $s' \preceq t'$ for some $s' \in [s]$ and $t' \in [t]$.

ensures that the matching is one-to-one and respects the order of time. The third condition, $f([s]) \leq s$, says that every high-level state is part of its associated low-level states; in other words, when we move from the low-level to the high-level, we do not add new information.

It is clear that the abstraction relation is reflexive and transitive, and hence that refinement, as the converse of abstraction, is also reflexive and transitive.

When one thinks of abstraction, it is natural to think of equivalence relations. However, there is an alternative way to define path abstraction, which is mathematically simpler since it relates paths directly without needing to pass through partitions.

6.2.3. DEFINITION (Abstraction function). Given a state space (S, \leq) and paths $l = (S_l, \preceq_l)$ and $h = (S_h, \preceq_h)$, $f : S_l \rightarrow S_h$ is an *abstraction function* with respect to (S, \leq) just in case it is

- surjective, for all $t \in S_h$, $t = f(s)$ for some $s \in S_l$
- order-preserving, if $s \preceq_l s'$ then $f(s) \preceq_h f(s')$, for all $s, s' \in S_l$
- and does not add new parts. $f(s) \leq s$ for all $s \in S_l$

It is straightforward to show that h is an abstraction of l in the sense of Definition 6.2.2 just in case there is an abstraction function from l to h .⁷

6.2.1 Examples of abstractions

We experience the world as continuous in time, but we often describe it in terms of discrete changes; saying, for example that the switch is flicked, then the light turns on, ignoring the perhaps infinitely many times when the charge was propagating through the wire. It is clearly useful to be able to abstract away from time's continuity and describe things discretely.

This abstraction from dense time to discrete time falls under the notion of abstraction we have just defined. For example, Figure 6.11 illustrates the switch and light in dense time.



Figure 6.11: The switch and light in dense time. ■ denotes that the switch is up, ■ that it is down, ■ that the light is off and ■ that it is on.

We can partition the dense intervals into discrete intervals, as in Figure 6.12.

This shows that the discrete path is an abstraction of the dense path, according to our definition of abstraction.

⁷Given \sim and $f : l/\sim \rightarrow h$ we take the abstraction function g with $g(s) = f([s])$; vice versa we do the same and take $s \sim s'$ just in case $g(s) = g(s')$.

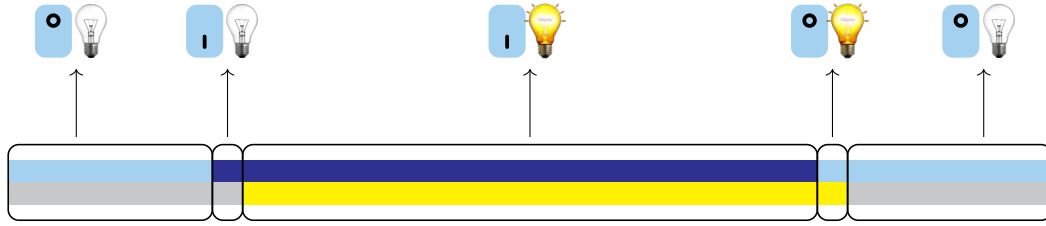


Figure 6.12: Abstraction discrete time from dense time.

We can also abstract away intermediate processes; for example, the electricity in the wire between the switch and light, as shown in Figure 6.13.⁸

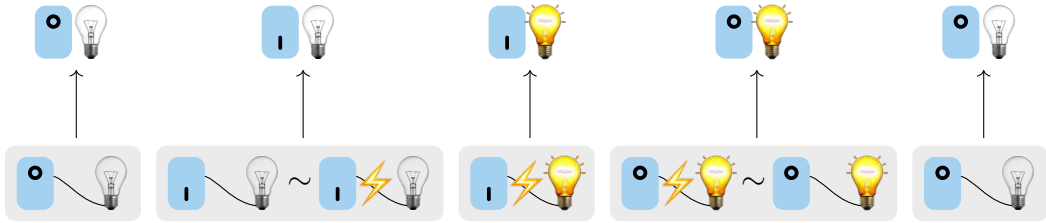


Figure 6.13: Abstracting away intermediate processes.

6.2.2 The interaction between abstraction and possibility

It is a general principle that the micro entails the macro. This principle can be expressed in many ways. Whenever a state is actual, all of its parts are actual too. Low-level descriptions entail their high-level counterparts: *x is scarlet* implies *x is red*; *x is a whale* implies *x is a mammal*. In biological taxonomy, if two organisms are in the same species, they are in the same genus. In statistical mechanics, whenever a system is in microstate s it is also in each of the macrostates associated with s . The presence of the token implies the presence of the type, and so on.

There is a corresponding principle for paths, which I take to be true: whenever a path is actual, so are its abstractions.

The micro-to-macro actuality principle.

If a path is actual, its abstractions are actual too.

To illustrate, consider the path with noise. If the world moved through the transitions described by the low-level path, it also moved through the transitions described by its abstraction.

⁸The example of abstracting away intermediate processes is also discussed by Rubenstein et al. (2017) in the context of structural equation models.

The micro-to-macro actuality principle implies, by contraposition, whenever a path is *not* actual, none of its refinements are too. If the world did not move through the transitions described by the high-level path, it also did not move through the transitions described by the low-level path.

This principle has consequences not only concerning when a path is actual, but modal consequences concerning when a path is possible. The micro-to-macro actuality principle implies the corresponding principle for possibility:

The micro-to-macro possibility principle.

If a path is nomically possible, its abstractions are nomically possible too.

Figure 6.14 illustrates these implications.

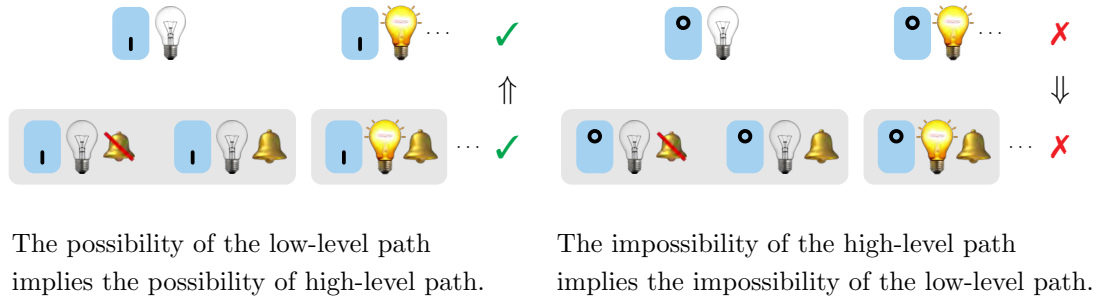


Figure 6.14

Let's how the former principle implies the latter. Suppose path p is nomically possible. This is equivalent to saying that it is nomically possible for p to be actual – that there is a nomically possible world w where p is actual. Then by the micro-to-macro actuality principle, every abstraction q of p is also actual at w . So there is a nomically possible world where q is actual; that is, q is nomically possible.

In symbols, this reasoning is an instance of $\Box(\varphi \rightarrow \psi) \rightarrow (\Diamond\varphi \rightarrow \Diamond\psi)$, a theorem of every normal modal logic. So for any abstraction q of p , and in every normal modal logic, we have the following implications.

$$\begin{aligned}
 & \vdash \text{actual}(p) \rightarrow \text{actual}(q) && \text{(micro-to-macro actuality principle)} \\
 \Rightarrow & \vdash \Box(\text{actual}(p) \rightarrow \text{actual}(q)) && \text{(necessitation)} \\
 \Rightarrow & \vdash \Diamond\text{actual}(p) \rightarrow \Diamond\text{actual}(q) && (\vdash \Box(\varphi \rightarrow \psi) \rightarrow (\Diamond\varphi \rightarrow \Diamond\psi))
 \end{aligned}$$

Since possible actuality is equivalent to possibility, $\Diamond\text{actual}(p)$ is equivalent to $p \in P$. What we have shown, then, is that assuming nomic possibility is a normal modality – which intuitively, it is – the micro-to-macro actuality principle implies the micro-to-macro possibility principle.

For any path p , let $\uparrow p$ be the set of its abstractions and $\downarrow p$ the set of its refinements (the notation is chosen to suggest that $\uparrow p$ contains the high-level

representations of p and $\downarrow p$ the low-level representations of p). And for any set of paths X , let $\uparrow X$ denote the set of abstractions of paths in X and $\downarrow X$ the set of refinements of paths in X .

$$\uparrow X = \bigcup_{p \in X} \uparrow p \qquad \downarrow X = \bigcup_{p \in X} \downarrow p$$

Since the abstraction and refinement relations are both reflexive and transitive, $\uparrow X$ is also the closure of X under abstraction and $\downarrow X$ the closure of X under refinements. That is, $\uparrow X$ is the smallest superset Y of X such that for every path in Y , its abstractions are also in Y ; similarly for $\downarrow X$ and refinements.

6.2.3 Partial models of nomic possibility

At the beginning of this chapter we took a model to specify which worlds are nomically possible and which are nomically impossible. The classification was exhaustive: a model decides for every world whether it is nomically possible (in P) or nomically impossible (not in P). In the partial setting, we instead specify which paths, rather than worlds, are nomically possible and which are impossible. We may wish to leave the status of some paths undecided. So we need two sets: a set of nomically possible paths and a set of nomically impossible paths. Let us denote these as P^+ and P^- , respectively.

6.2.4. DEFINITION (Partial model). A *partial model* (of the nomic possibilities) is a pair (P^+, P^-) where P^+ and P^- are sets of paths.

As we have seen, partial models of nomic possibility have consequences beyond the paths in P^+ and P^- . A path is nomically possible according to $P = (P^+, P^-)$ just in case it is the abstraction of a path in P^+ , and a path is nomically impossible according to P just in case it is the refinement of a path in P^- . That is, for any partial model P , $\uparrow P^+$ is the set of paths that are possible according to P and $\downarrow P^-$ is the set of paths that are impossible according to P .

From this we define what it means for two partial models to be incompatible. They are incompatible just in case there is a path that is possible according to one and impossible according to the other.

6.2.5. DEFINITION (Compatibility and consistency of partial models). Partial models $P = (P^+, P^-)$ and $Q = (Q^+, Q^-)$ are *compatible* just in case

$$\uparrow P^+ \cap \downarrow Q^- = \emptyset \qquad \text{and} \qquad \downarrow P^- \cap \uparrow Q^+ = \emptyset.$$

P is *consistent* just in case $\uparrow P^+ \cap \downarrow P^- = \emptyset$. Equivalently, P is consistent just in case it is compatible with itself.

We can also define an abstraction relation over partial models. Intuitively, a partial model P is an abstraction of a partial model Q just in case every path that is possible according to P is possible according to Q and every path that is impossible according to P is impossible according to Q .

6.2.6. DEFINITION (Partial model abstraction and refinement). For any partial models $P = (P^+, P^-)$ and $Q = (Q^+, Q^-)$, P is an *abstraction* of Q just in case

$$\uparrow P^+ \subseteq \uparrow Q^+ \quad \text{and} \quad \downarrow P^- \subseteq \downarrow Q^-.$$

We say Q is a *refinement* of P just in case P is an abstraction of Q .

Our definition of partial model refinement is analogous to the definition of refinement from three-valued logic. In that context, a valuation is a partial function from atomic sentences to the values *true* and *false*, and a valuation v *refines* a valuation u just in case every atomic sentence that is true in u is true in v and every atomic sentence that is false in u is false in v .⁹

Since $\uparrow Q^+$ is closed under abstraction, $\uparrow P^+ \subseteq \uparrow Q^+$ is equivalent to $P^+ \subseteq \uparrow Q^+$; likewise, since $\downarrow Q^-$ is closed under refinement, $\downarrow P^- \subseteq \downarrow Q^-$ is equivalent to $P^- \subseteq \downarrow Q^-$. So we can give the following simpler understanding of model abstraction:

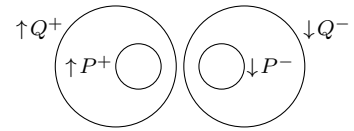
6.2.7. FACT. $P = (P^+, P^-)$ is an abstraction of $Q = (Q^+, Q^-)$ just in case

- every path in P^+ is an abstraction of a path in Q^+ , and
- every path in P^- is a refinement of a path in Q^- .

There is an interaction between consistency and abstraction:

6.2.8. FACT. If a partial model is consistent, its abstractions are consistent too. Equivalently, if a partial model is inconsistent, its refinements are inconsistent.

In other words, consistency is preserved ‘upward’ and inconsistency is preserved ‘downward’. This follows from the fact, illustrated on the right, for any sets A and B and subsets $A' \subseteq A$ and $B' \subseteq B$, if A and B are disjoint then so are A' and B' .



There is also the following interaction between compatibility and abstraction:

6.2.9. FACT. If a partial model is consistent, it is compatible with all of its abstractions. And if it is inconsistent, it is incompatible with all of its refinements.

⁹Refinement sometimes goes by other names in the literature on three-valued logic: Fine (1975b:268) writes that v *extends* u , Priest (2008:133) that v is a *resolution* of u .

Here is a final remark before we move on. Let us say that a partial model $P = (P^+, P^-)$ is *exhaustive* just in case every path is in $\uparrow P^+$ or $\downarrow P^-$. That is, a partial model is exhaustive just in case it decides the status of every path: every path is either possible according to P or impossible according to P . It follows immediately that for consistent and exhaustive models, compatibility and refinement coincide.

6.2.10. FACT. For any partial models $P = (P^+, P^-)$ and $Q = (Q^+, Q^-)$, if P is consistent and exhaustive then P is compatible with Q just in case P refines Q .

To illustrate our definition of model abstraction, recall the model of the switch and light from Figure 6.7, where every infinite path through the figure that follows the arrows is nomically possible, and every infinite path through the figure that does not follow the arrows is nomically impossible. Similarly, Figure 6.16 depicts the nomic possibilities and impossibilities for the switch, light and noise. The noise is independent of the switch and light, in the sense that it can change freely. As before, an infinite path through the Figure is nomically possible just in case it always follows the arrows.

Let $SL = (SL^+, SL^-)$ be the possibilities and impossibilities, respectively, of the switch-and-light model in Figure 6.15, and $SLN = (SLN^+, SLN^-)$ those of the switch-light-and-noise model in Figure 6.16. Notice that the two figures contain different paths: the sets of possibilities SL^+ and SLN^+ are disjoint, similarly for the impossibilities SL^- and SLN^- . This illustrates the importance of closing the set of possibilities under abstraction and the set of impossibilities under refinements. Every possibility for SL is a possibility for SLN , since every path in SL^+ is an abstraction of a path in SLN^+ . Conversely, every impossibility for SLN is an impossibility for SL , since every path in SLN^- is a refinement of a path in SL^- . So (SL^+, SLN^-) is an abstraction of (SLN^+, SL^-) .

Now, there is a clear sense in which the switch-light-noise model ‘entails’ the switch-light model. Whenever the switch-light-noise model is a correct description of the nomic possibilities, the switch-light model is too.

However, our definition of partial model abstraction in Definition 6.2.6 does not capture this: SL is not an abstraction of SLN , since $SL^- \not\subseteq \downarrow SLN^-$. For example, the path on the right is in SL^- , but since SL describes paths at a higher level than SLN , this path is not the refinement of any path in SLN^- .



The problem, generally speaking, is that model abstraction is represented by inclusion of both possibilities and impossibilities, and as we have seen, possibility and impossibility have different monotonicity properties: possibility is upwards closed (closed under abstraction) while impossibility is downwards closed (closed under refinement). Requiring inclusion of both possibility and impossibility results in non-monotonicity. The result is that, given a model containing high-level

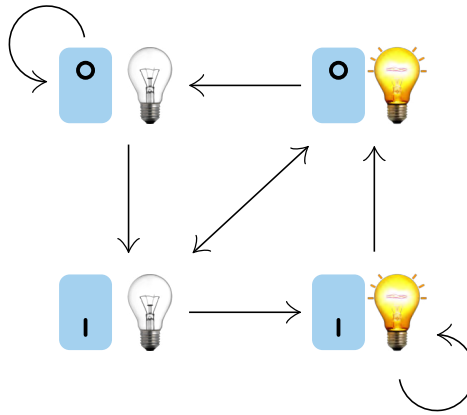


Figure 6.15: A world is nomically possible just in case it is an infinite directed path through the figure above.

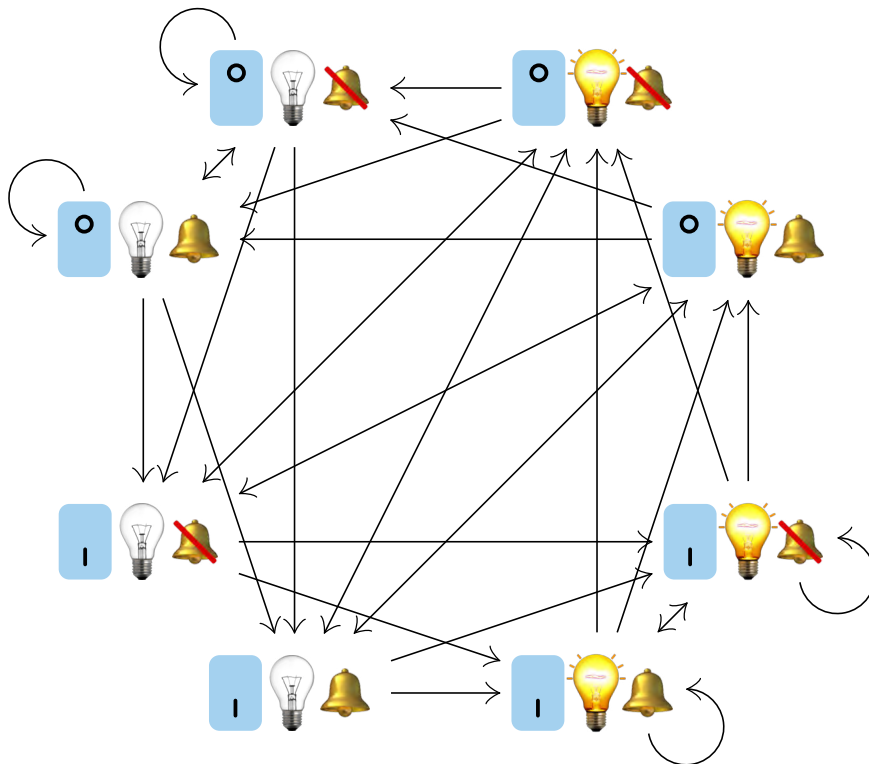


Figure 6.16: The signal and the noise.

paths and model containing low-level paths, in general the high-level model will not be an abstraction of the low-level model.

We can rectify this situation by taking a global perspective, understanding abstraction with respect to worlds rather than paths. Recall that a path is a linear order of situations, while a world is a linear order of maximal situations. In our model construction from the beginning of this chapter a model specified which worlds are nomically possible. We can extend this to a classification over all paths in a straightforward way. Given the set of nomically possible worlds, a path is nomically possible just in case it is the abstraction of a nomically possible world. Where $WM \subseteq W$ is a set of worlds, let us call WM a *world model*. Every world model WM generates a partial model $(\uparrow WM, \overline{\uparrow WM})$, where $\overline{\uparrow WM}$ is the set of paths not in $\uparrow WM$.

6.2.11. DEFINITION. For any partial model $P = (P^+, P^-)$ and world model WM , P is an abstraction of WM just in case P is an abstraction of $(\uparrow WM, \overline{\uparrow WM})$. And P is compatible with WM just in case it is compatible with $(\uparrow WM, \overline{\uparrow WM})$.

As usual, we say WM *refines* P just in case P is an abstraction of WM .

Equivalently, P is an abstraction of $(\uparrow WM, \overline{\uparrow WM})$ just in case every path in P^+ is the abstraction of a world in WM , and no path in P^- is the abstraction of a world in WM .

We can then define a notion of entailment between partial models as follows.

6.2.12. DEFINITION. For any partial models P and Q , P *entails* Q just in case every world model that refines P refines Q .

That is, P entails Q just in case for any world model WM , if every path in P^+ is the abstraction of a world in WM and no path in P^- is, then every path in Q^+ is the abstraction of a world in WM and no path in Q^- is.

We may equivalently define model entailment in terms of compatibility:

6.2.13. FACT. For any partial models P and Q , P entails Q just in case every world model compatible with P is compatible with Q .

This holds since for any world model WM , the partial model generated by WM , $(\uparrow WM, \overline{\uparrow WM})$, is consistent and exhaustive, so by Fact 6.2.10, any world model refines a partial model P just in case it is compatible with P .

Our definition of partial model entailment bears an obvious resemblance to the standard notion between propositions, whereby one proposition entails the other just in case in every world where the first is true, the second is also true. Indeed, it is natural to give the following definition.

6.2.14. DEFINITION (Truth at a world model). A partial model P is *true* at a world model WM just in case P is an abstraction of WM .

That is, P is true at WM just in case every path that is possible according to P is possible according to WM , and every path that is impossible according to P is impossible according to WM . Then P entails Q just in case every world model where P is true, Q is true. This formalises what it means for the truth of one model to imply the truth of another. In other words, P entails Q just in case whenever P is a true description of the nomic possibilities, Q is too. For example:

6.2.15. FACT. The switch-light-noise model entails the switch-light model.

PROOF. Pick any world model WM that refines the switch-light-and-noise model. Then $SLN^+ \subseteq \uparrow WM$ and $SLN^- \subseteq \overline{\uparrow WM}$. To show that WM refines the switch-and-light model, we have to show (i) $SL^+ \subseteq \uparrow WM$ and (ii) $SL^- \subseteq \overline{\uparrow WM}$.

(i) Note that every possible path in the switch and light model is an abstraction of a possible path in the switch-light-and-noise model: $SL^+ \subseteq \uparrow SLN^+$. And since $\uparrow WM$ is closed under abstraction, $SLN^+ \subseteq \uparrow WM$ is equivalent to $\uparrow SLN^+ \subseteq \uparrow WM$. Thus $SL^+ \subseteq \uparrow SLN^+ \subseteq \uparrow WM$.

(ii) This follows from the fact that (*) every world that refines a path in SL^- refines a path in SLN^- . To show that $SL^- \subseteq \overline{\uparrow WM}$, pick any path $p \in SL^-$ and suppose for reductio that $p \in \uparrow WM$. Then some world w refines p , so by (*), w refines some path in SLN^- . So $q \in \uparrow WM$. But since $SLN^- \subseteq \overline{\uparrow WM}$, $q \in \overline{\uparrow WM}$, contradicting $q \in \uparrow WM$. Hence $p \notin \uparrow WM$, that is, $p \in \overline{\uparrow WM}$, as required. \square

As we saw above, the switch-light model is not an abstraction of the switch-light-noise model. This is because the switch-light model operates at a higher level than the model with noise: the paths of the higher model are not refinements of the paths of the lower model. Abstraction therefore misses out on an important relationship between the two models. Our definition of entailment solves this issue. The key to the result that the switch-light-noise model entails the switch-light model is that every world that refines a path in the switch-light model also refines a path in the switch-light-noise model. This follows from the fact that the switch-light-noise model describes all the possible states of the switch, light and noise. As worlds are linear orders of maximal situations, every situation in a world w will decide the status of the switch, light and noise, and w will therefore be the refinement of some path in the switch-light-noise model.

Conversely, the switch-light model does not entail the switch-light noise model. This is because the latter asserts more possibilities than the former. For example, a world model where it is impossible for the noise to turn on is compatible with the switch-light model but not with the switch-light-noise model.

6.2.4 A global view

At the beginning of this section we mentioned that logic, linguistics and philosophy often take what we called a ‘global’ view. For example, in possible worlds

semantics one typically begins with a set of worlds – a maximally inclusive space of possibilities – and understands meanings in terms of this logical space. We mentioned that one advantage of this view is that the single all-encompassing model serves as a common forum in which all meanings can be compared.

The goal of this section is to show that we can take a global perspective on the present framework, and thereby inherit its benefits. In standard possible worlds semantics worlds are taken to be maximal and consistent points of evaluation. The analogue in our case is a consistent and complete specification of the nomic possibilities – what we will call a *full* model.

6.2.16. DEFINITION (Full model). A full model is a pair $M = (M^+, M^-)$, where M^+ and M^- are sets of paths, that is exhaustive and consistent: every path is in $\uparrow M^+$ or $\downarrow M^-$ and no path is in both.

A full model is a kind of partial model, one that is consistent and determines for each path whether it is nomically possible or impossible.

In possible worlds semantics one represents the meaning of a declarative sentence as the set of worlds where it is true. Similarly, we can think of a partial model as a kind of assertion: it asserts that some paths are nomically possible and others are nomically impossible. What does it mean for a partial model to be true at a full model? There is a natural answer: a partial model P is true at a full model M just in case it agrees with M on the nomic possibilities and impossibilities: every path that is possible according to P is possible according to M , and every path that is impossible according to P is impossible according to M ; in other words, M refines P . And just as we think of the proposition expressed by an assertion as the set of worlds where the assertion is true, we may think of the proposition expressed by a partial model as the set of full models where it is true: for any partial model P , let $|P|$ be the set of full models that refine P . Then P is true in M just in case $M \in |P|$.

6.2.17. DEFINITION (Truth at a full model). A partial model P is true at a full model M just in case M refines P . Let $|P|$ denote the set of full models where P is true.

We noted above that refinement can be viewed as a kind of entailment. In possible worlds semantics one defines entailment in terms of inclusion: A entails B just in case every world where A is true, B is true too. One can treat refinement in an analogous way, as shown by the following.

6.2.18. PROPOSITION. For any partial models P and Q ,

$$|P| \subseteq |Q| \quad \text{if and only if} \quad P \text{ refines } Q, \text{ or } P \text{ is inconsistent.}$$

PROOF. Let $P = (P^+, P^-)$ and $Q = (Q^+, Q^-)$ be partial models. (\Rightarrow) Suppose $|P| \subseteq |Q|$ and that P is consistent. Consider $M = (\uparrow P^+, \overline{\uparrow P^+})$ and $N = (\downarrow \overline{P^-}, \downarrow$

P^-), where \overline{X} is the set of paths not in X . Clearly, M and N are full models (consistent and exhaustive). We show that M refines P , that is, $\uparrow P^+ \subseteq \uparrow\uparrow P^+$ and $\downarrow P^- \subseteq \downarrow(\uparrow\overline{P^+})$. The first inclusion is immediate. To show the second, pick any $p \in \downarrow P^-$, i.e. p refines some $p' \in P^-$. Since P is consistent, $p' \notin P^+$, i.e. $p' \in \overline{P^+}$, so $p \in \downarrow(\uparrow\overline{P^+})$, as required. One similarly shows that N refines P . Hence $M, N \in |P|$, so $M, N \in |Q|$. Since M refines Q , $\uparrow Q^+ \subseteq \uparrow\uparrow P^+ = \uparrow P^+$ and since N refines Q , $\downarrow Q^- \subseteq \downarrow\downarrow P^- = \downarrow P^-$. Hence P refines Q .

(\Leftarrow) Suppose that if P is consistent, P refines Q . If P is inconsistent, by Fact 6.2.8 all of its refinements are inconsistent too. Since full models are consistent by definition, $|P| = \emptyset$, so vacuously $|P| \subseteq |Q|$. Now suppose P is consistent. Then P refines Q . Since refinement is transitive, every model that refines P refines Q . A fortiori, every full model that refines P refines Q : $|P| \subseteq |Q|$. \square



Proposition 6.2.18 builds a bridge between a local perspective, which compares two partial models directly using refinement, and a global perspective, which compares them in terms of the set of full models.

It also shows that partial models inhabit the same mathematical structure as propositions in possible worlds semantics, that of a Boolean algebra. To see this, let us say that two partial models P and Q are equivalent just in case, if they are consistent, then they agree on which paths are nomically possible and which are nomically impossible: $\uparrow P^+ = \uparrow Q^+$ and $\downarrow P^- = \downarrow Q^-$. Then Proposition 6.2.18 shows that the class of partial models, identified up to equivalence, forms a Boolean algebra with the order given by $P \leq Q$ just in case if P is consistent, P refines Q . The top element in the algebra represents the trivial model where P^+ and P^- are both empty, (\emptyset, \emptyset) , which does not decide the status of any paths, while the bottom element represents the inconsistent models.

In our model construction from beginning of this chapter we took a model to specify which *worlds* are nomically possible and which are nomically impossible. This can be naturally extended to a full model, determining for every *path* whether it is possible or impossible: we say that a path is nomically possible just in case it is the abstraction of a nomically possible world. In this way, given the set P of nomically possible worlds, P generates a full model $(\uparrow P, \uparrow\overline{P})$, where $\uparrow\overline{P}$ is the set of paths not in $\uparrow P$, and we may say that a partial model Q is true according to P just in case it is true according to the full model generated by P .

This completes our demonstration that we can recreate the global view popular in logic, linguistics and philosophy within the present framework.

6.3 Representing causal asymmetry

It may be surprising to see a state like   in a nomically possible world, where the switch has been flicked down but the light is still off. This state violates the

claim that the light is off if and only if the switch is up. But that claim is only true when the system is in equilibrium. A state where the switch has been flicked and the light has not yet changed is perfectly possible. It takes some time for the light to turn on after the switch has been flicked. There is nothing unlawful about non-equilibrium states. We experience such states all the time, even if only briefly. Such states play an essential role in representing causal asymmetry. They show that a change in the switch must lead to a change in the light, but a change in the light need not lead to a change in the switch.

To illustrate, consider a model where the roles of the switch and light are reversed: a change in the light must lead to a change in the switch, but a change in the switch need not lead to a change in the light. What results is a very different model, one that does not represent the actual behaviour of the switch and light at all. This is illustrated Figure 6.17.

But now imagine we ignored the states where the switch has been flicked and the light has yet to change; in other words, suppose our model contains only states where $L \leftrightarrow S$ is true: the light is on if and only if the switch is up. The difference between the correct model and the incorrect reversed model vanishes.

This illustrates how states where the cause has occurred but the effect is yet to occur play an crucial role in representing causal asymmetry. More generally, representing causal asymmetry makes essential appeal to the asymmetry of time.

6.3.1 The asymmetry of *cause*

Now, this by itself does not derive the fact that the verb *cause* is asymmetric, in the sense that whenever C *cause* E is true, E *cause* C is false. For this we need to look to the meaning of *cause*. On our account, the asymmetry of *cause* follows from two facts concerning the notion of production used by *cause* (namely, using proper chains; see section 5.4). Causation entails production_{proper}, and production_{proper} is asymmetric: whenever C produced_{proper} E , E did not produce_{proper} C , so we have the following implications.

$$C \text{ cause } E \Rightarrow C \text{ produce}_{\text{proper}} E \Rightarrow \neg(E \text{ produce}_{\text{proper}} C) \Rightarrow \neg(E \text{ cause } C)$$

As we discussed in chapter 3, proper production is asymmetric since it requires chains that move forward in time (unlike the notion of production used by *because*, which is improper). Thus C produce_{proper} E implies that C occurred before E . Then by the asymmetry of time, E did not occur before C , so E produce_{proper} C is false.

6.3.2 Alleged cases of simultaneous causation

The present framework uses the asymmetry of time to model causal asymmetry. Such an account faces alleged cases of simultaneous causation (Gasking 1955,

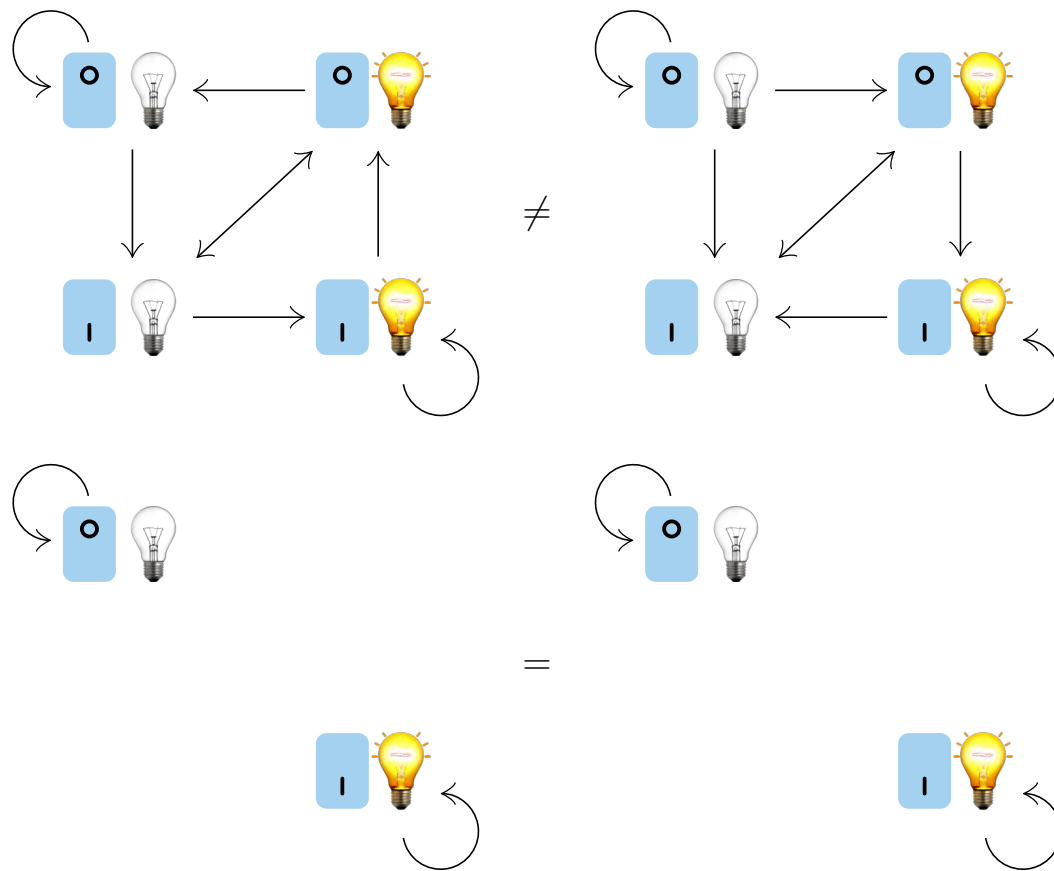


Figure 6.17: If we restrict to states where the light is on if and only if the switch is up, the original and reversed model become equivalent – the causal asymmetry between the switch and the light disappears.

Taylor 1966, Brand 1979, Brody 1980). Taylor (1966) considers a trailer hitched to a car, arguing that “They move together, and in no sense is the motion of one temporally followed by the motion of the other.” Brand (1979:272–273) observes that Jack going down on a seesaw causes Mary to go up at the same time. Brody (1980) imagines compressing a sealed container, which increases the pressure inside. The decrease in volume causes the increase in pressure, but according to Brody the changes happen at the same time.

Kline (1980) responds that these are not cases of simultaneous causation, which I agree with. A delay between cause and effect becomes apparent once we inspect the physical make up of the scenarios. For example, the car and trailer are held together by electromagnetic forces, which cannot propagate faster than the speed of light, leading to a delay between the car moving and the trailer following. If the trailer were instead pulling the car, the forces would propagate

in the opposite direction. The same applies to the seesaw. Brand imagines the seesaw to be a perfectly rigid body, which is impossible. About the pressure example, Kline writes,

we tend to think that the relations between [volume] and [pressure] are instantaneous. However, ... the gas laws are intended as descriptions of gases in *equilibrium*. They are not to be taken as describing processes. ... Once a change in the volume occurs it is some time until the gas reaches equilibrium and is correctly described by [the ideal gas law, $PV = NrT$].

We can begin to see our way clear of this problem if we consider the behavior of gases from a molecular point of view. ... Pressure is defined as force per unit area. ... The pressure will increase at different points on the container at different times. This is one reason why the gas laws are restricted to equilibrium states – a point after the change when the pressure, temperature, etc. are uniform.

(Kline 1980:296–97)

The increase in pressure begins where the container is compressed and propagates in time.

One may reply that, even though simultaneous causation is not possible given the current laws of physics, it is possible in other worlds with different laws, and since we would like our analysis of causation to apply no matter what the laws happen to be, our analysis should be compatible with simultaneous causation.

This reply assumes that in a world where the laws are different, we would still accept the same causal claims; say, that change in volume caused the change in pressure and not vice versa. That is not guaranteed. It is universally agreed that causal relations depend on the laws. The present account of the meaning of *cause* and *because* reflects this since their truth requires sufficiency, which depends on the nomic possibilities. We do not have a reliable way of knowing what causal claims we would accept in a world with different laws – say, in a world with perfectly rigid bodies – since all of our experience comes from objects governed by the actual laws. We may ask each other to imagine a perfectly rigid body, but when imagining the scenario are likely importing assumptions from our intuitive physics, one that is steeped in our experience with the actual laws.

6.3.3 Kim's cases

Here is a second, more challenging alleged case of simultaneous causation, inspired by Kim (1973, 1974).

- (1)
 - a. The birth of my niece caused me to become an uncle.
 - b. Socrates dying caused Xanthippe to become a widow.

Kim (1973) states that “My becoming an uncle was determined by, was dependent on, the birth of the child, but was not a causal effect of it”. Kim (1974:49) even argues that nothing at all is a cause of Xanthippe becoming a widow. Sartorio (2006) agrees that Xanthippe’s widowhood was not caused by Socrates’ death.

That being said, to my ear the sentences in (1) sound quite good. Regardless whether one thinks this should count as a case of causation, we would like to account for the acceptability of (1).

A second datum is that there is a stark contrast between (1) and the reverse claims, which are clearly unacceptable.

- (2) a. Me becoming an uncle caused the birth of my niece.
- b. Xanthippe becoming a widow caused Socrates to die.

At first glance it seems hard to account for the contrast between (1) and (2) using temporal asymmetry. For it is plausible to say that Socrates died at the exact same time Xanthippe became a widow. How then can we account for the acceptability of (1) and unacceptability of (2) on the present account?

Here we will only focus on (1b) and (2b). Our account can be easily extended to account for the contrast between (1a) and (2a).

I believe we can account for them when we attend to the temporal structure of events. Thankfully there has been a great deal of work on this topic in semantics. To analyse these cases in particular, we have to clarify the temporal semantics of the verbs *die* and *become*. According to Vendler (1957), *die* and *become* are both achievements (yes, in linguistics to die is classed as an achievement!). A prominent idea is that achievements denote culminations – instantaneous endpoints – of events.¹⁰ For example, *die* denotes the culmination of the event of dying.

Piñón (1997) argues that achievements may also denote beginnings, as shown by verbs such as *reach* (as in *reach the summit*) and *recognize*: “Take *recognize*: it plausibly denotes beginnings of states of recognizing. Accordingly, if Anita suddenly recognizes Peter, then there is a beginning of a state in which she recognizes him” (Piñón 1997:277). I agree with Piñón’s analysis. The verb *become* is a clear case of an achievement that denotes the beginning of a state: *become a widow* denotes the beginning of the state of being a widow.

It is standardly assumed that achievements denote instants. Vendler (1957) writes that achievements “occur at a single moment”. Following Piñón (1997), let us therefore assume that beginning achievements (such as *become a widow*) occur at the left boundary of their associated eventuality and ending achievements (such as *die*) occur at the right boundary of their associated eventuality. Formally, for any interval (i.e. dense set of time points) *I*, we define the *left boundary* of *I* to be its greatest lower bound: the latest time that is earlier than or equal to every time in *I*. Similarly, we define the *right boundary* of *I* to be its least upper bound:

¹⁰See e.g. Moens and Steedman (1988), Binnick (1991:195) and Kamp and Reyle (1993:§5.3.2).

the earliest time that is later than or equal to every time in I . This implies that for any intervals $[t, t']$, $[t, t')$, $(t, t']$ and (t, t') , their left boundary is t and their right boundary is t' , which seems reasonable.

There are therefore two relevant eventualities involved in analysing the sentence, *Socrates dying caused Xanthippe to become a widow*: the event of Socrates dying, and the state of Xanthippe being a widow. Our task now is to understand the temporal relation between them. It is plausible to assume that they do not overlap. There is no time at which Socrates is in the process of dying and Xanthippe is a widow. There also does not appear to be any time strictly between these eventualities. Once Socrates is dead Xanthippe is immediately a widow.

These two observations – no temporal overlap between the two eventualities but also no time strictly between them – leave two possibilities: where d is the runtime of Socrates' death and w the runtime of the state of Xanthippe's widowhood, we can have (i) d is closed to the right and w is open on the left, or (ii) d is open to the right and w is closed to the left. In either case we predict that Socrates died at the same time that Xanthippe became a widow. This is the intuitively correct result. What is perhaps surprising is that holds even though the two associated eventualities (the event of Socrates' dying and the state of Xanthippe's widowhood) in fact do not overlap in time. There is also a temporal asymmetry between the two: the event of Socrates' dying precedes the state of Xanthippe being a widow. This is illustrated in Figure 6.18 with case (i).

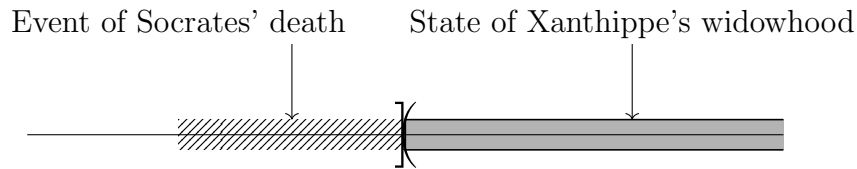


Figure 6.18

We will exploit this temporal asymmetry to account for the acceptability of (1) and unacceptability of (2). C *cause* E entails that C produced_{cause} E . Given our analysis of production from chapter 3, C produced_{cause} E just in case there is a proper chain of proposition–time pairs such that each proposition not being true at that time is sufficient for the next proposition to not be true at its time. To check sufficiency we remove the part of the world the event the proposition is about while holding the past of that event fixed.

While Socrates died at exactly the same time Xanthippe became a widow, the phrase *Socrates' death* is about the event of Socrates' death. When we remove that event, we fix the state of the world prior to that event (see Figure 6.19).

The past we fix when evaluating whether Socrates dying produced Xanthippe to become a widow allows Socrates to die at different times, and therefore also for Xanthippe to become a widow at different times than she in fact did. The dependence required for production is established, so Socrates dying produced

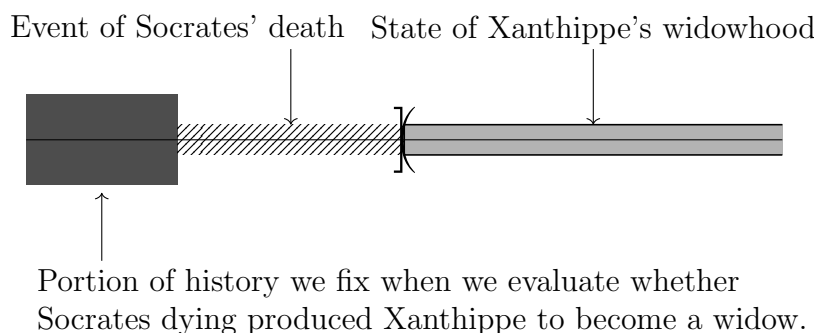


Figure 6.19

Xanthippe to become a widow. Given that this holds for every way in which Socrates could die (given the circumstances), Socrates dying was sufficient for Xanthippe to become a widow: the positive condition is satisfied. And since Xanthippe would not have become a widow if Socrates had not died, the negative condition is also satisfied. We therefore predict the causal claim in (1b) to be true.

In contrast, *Xanthippe became a widow* is about a state that begins at exact time when the event of Socrates' death ends. Since production requires us to fix the past of the event the proposition is about, we fix fact that the event of Socrates' death occurred at the time that it did (see Figure 6.20).

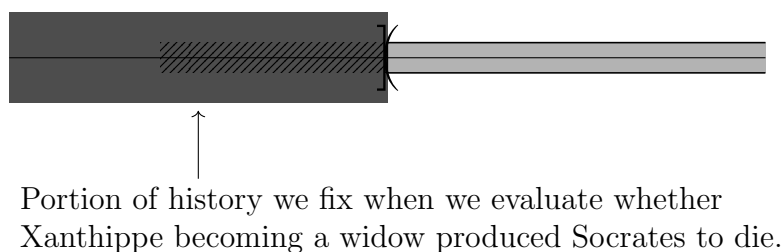


Figure 6.20

This guarantees that Socrates still dies at time when he actually did (since in any logically possible world, the right boundary of the event of Socrates' death will be the same). The dependence required for production fails, so Xanthippe becoming a widow did not produce Socrates to die. We therefore predict (2b) to be false, as desired.

6.4 The concept of nomic possibility

A common idea is that causation and laws of nature are intimately connected. This is most obvious on regularity approaches to causation, where the laws of

nature are taken to be a particular kind of regularity (as e.g. Mill 1843 proposed), but this idea is to some extent part of every theory of causation. One might therefore wonder what notion of laws is assumed in the present framework.

We follow the strategy Lewis proposed when faced with a notion of comparable difficulty to that of laws of nature: meaning. “To say what a meaning is,” Lewis writes, “we may first ask what a meaning does, and then find something that does that” (Lewis 1970a:22). Similarly, to say what a law of nature is, we may first ask what a law of nature does, and then find something that does that.

What a law does, at its most basic, is categorise behaviour. In the simplest case there are two categories: legal and illegal. We may call two laws (or sets of laws) equivalent just in case they agree on what behaviour is legal and what behaviour is illegal.¹¹ The simplest possible formalisation of the laws then, is a pair (L^+, L^-) where L^+ is the set of things compatible with the laws and L^- is the set of things incompatible with them.

Notice that our formalisation of the laws of nature is purely extensional. It does not require laws to be written in some language or to be axiomatised by some deductive system. We will return to this point in section 6.4.1.

A law is consistent when there is no overlap: no behaviour is legal and illegal with respect to it. A law is exhaustive when there are no gaps: every behaviour is either legal or illegal with respect to it. If a law is consistent and exhaustive, and it is clear what the domain of the law is (i.e. what things it categorises, such as human behaviour or natural phenomena), then the formal representation simplifies from the pair (L^+, L^-) to either element by itself, since given the domain each may be reconstructed from the other (e.g. L^- is the set of things in the law’s domain not in L).

What is the domain of the laws of nature? That is, what kind of thing do laws of nature categorise as legal or illegal? We may take it to be all logically or metaphysically possible phenomena. But this is more than we need.

We will instead take nomic possibility to be a property of all logically and metaphysically possible worlds. Why worlds? Why not something smaller – proper parts of worlds, such as states of affairs, events, or the state of the world at a moment in time? We do not take nomic possibility to be a property of proper parts of worlds because whether something is nomically possible depends not only on itself but also on its surroundings. Firstly, nomic possibility is sensitive to the spatial environment. It may be nomically possible for Alice to throw a ball five

¹¹Of course, for laws in real life this extensional view is too simple. Laws do many other things: give definitions, make political statements, express attitudes and so on. Two laws expressed in real life may agree on which behaviours are legal and illegal but still disagree on their political effect. The point applies not only to laws of politics but also laws of physics. Two physical theories may be equivalent in the sense that they make the same predictions but differ in emphasis, ease of use or insight (such as Newtonian, Lagrangian and Hamiltonian formulations of classical mechanics). Our question here is whether the semantics of causal claims requires this enriched notion of law, or merely the classification into legal and illegal.

meters in the air, but not if she is in a cave. Secondly, nomic possibility is sensitive to the temporal environment. It is not nomically possible for there to be no ball in Alice’s vicinity at one moment, but one to spontaneously appear in her hand the next (what is nomically possible right now depends on the past). If we only recorded whether the proper parts of a world are nomically possible, without also recording whether the world as a whole is, we would not know whether it is nomically possible for the parts to coexist as a whole.

We assume that the laws of nature are consistent and exhaustive. This is equivalent to saying that a world is nomically impossible just in case it is not nomically possible. Formally, then, nomic possibility is a simply property of worlds. So nomic possibility can be formally represented as a set of worlds – those that are taken to be nomically possible.¹²

One may think that to give the truth conditions of causal claims, we need to go further by analysing what laws of nature are. However, this is unnecessary if the semantics of causal claims only requires an extensional view of laws, in the sense that the semantics does not care how the laws are formulated, but only what they do; that is, if the semantics only cares about what worlds are possible and what worlds are impossible according to them. The most direct way to show that the semantics of causal claims only requires an extensional view of laws is to provide such a semantics and show its empirical adequacy. This we have already tried to do: we reduced the semantics of *cause* and *because* to sufficiency, and our definition of sufficiency only required a classification of worlds into lawful and unlawful.

6.4.1 Comparison with proof-theoretic views of laws of nature

Standardly, something is nomically possible just in case it is compatible with the laws of nature – *nomos* being the Greek for ‘law’. One could try to formally represent nomic possibility by analysing the components of this statement, i.e. what makes something a law of nature, and what it is to be compatible with it. This is, of course, an extremely challenging task (for an overview of the issues involved see Carroll 2020). To give a taste of the difficulty, consider Ramsey’s proposal that causal laws are

“consequences of those propositions which we should take as axioms if we knew everything and organized it as simply as possible in a deductive system”

(Ramsey 1929b:150)

or Lewis’s proposal that

¹²We could of course have taken the set of nomically impossible worlds as primitive and defined the nomically possible ones as those that are not nomically impossible.

“a contingent generalisation is a *law of nature* if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength.”

(Lewis 1973b:73)

These analyses raise tricky questions, such as how to measure simplicity, in what languages the laws are written, and what deductive system(s) to use. Nonetheless, this view of laws is often praised as faithfully representing the concept of law that physicists actually work with; for example, physicists do write their laws in a particular language and do aim for simple, general laws. So let us grant, for the sake of argument, that we have answers to all of these questions: we have an ideal language in which to write the laws, we know everything, physics is complete, and we know exactly which deductive systems organise our knowledge as simply as possible. Given these generous concessions, Ramsey and Lewis’ conception of laws of nature comes down to the following claim.

The proof-theoretic view of laws of nature. There is a deductive system (however ideal) such that a sentence is a law of nature just in case it is a theorem of that system.¹³

Here is an argument showing the limits of this view. Some problems in physics have been shown to be incomplete, in the sense that their answer is independent of the axioms of mathematical theory in which they are expressed (see Barrow 2011 for discussion). One shows this by replicating Gödel’s first incompleteness theorem in theories of physics rather than arithmetic. For example, da Costa and Doria (1991) prove a version of Gödel’s incompleteness theorem within Hamiltonian mechanics.

Incompleteness results in physics are not that surprising. The level of mathematical strength required for Gödel’s incompleteness theorems is rather weak (Robinson arithmetic is enough, which does not even have induction; see Tarski, Mostowski, and Robinson 1953), and is easily exceeded by much of the mathematics applied in theories of physics.

Given the incompleteness of physical theories, any deductive system with sufficient strength will fail to prove some sentences in the language of that theory which are in fact true according to it. This includes, one assumes, the ideal deductive systems that Ramsey and Lewis dream of. The proof-theoretic view of laws is too restrictive: it ends up not counting some sentences as laws that intuitively ought to count as laws.¹⁴

¹³Alternatively, we may express the proof-theoretic view of laws as saying that there are some deductive systems such that a sentence is a law of nature just in case it is a theorem of *all* of them.

¹⁴There is a separate but related issue, one that does not challenge the proof-theoretic view of laws but does challenge views of laws based on computability, such as those who seek to represent the world as some kind of computer. The issue is that some naturally occurring

A potential response to incompleteness results in physics is that they result from the idealisation of mathematics. For example, Barrow (2011) writes:

Another possible way of evading Gödel’s theorem is if the physical world only makes use of the decidable part of mathematics. We know that mathematics is an infinite sea of possible structures. Only some of those structures and patterns appear to find existence and application in the physical world. It may be that they are all from the subset of decidable truths.

This stipulation would save the proof-theoretic view of laws, but at the cost of having our view of the laws of nature dictate physics. This is a gratuitous commitment. Our view of the laws of nature should work regardless what the laws actually are. In particular, it should apply to worlds where the mathematical idealisations that physicists make happen to be true.

In contrast, the view of the laws adopted here – namely, a classification of worlds into those that satisfy the laws and those that do not – avoids this problem. We have opted for a semantic rather than proof-theoretic analysis of the laws of nature. This does not dictate what kind of mathematics physics is allowed to apply, since we do not assume the existence of a deductive system that is complete with respect to the set of true laws.

Indeed, our view is strictly more general than the proof-theoretic view. For suppose one has an ideal theory of physics (or set of theories) that Ramsey and Lewis envisage. To apply the theory, we also require a way to interpret it – that is, a semantics. This generates a classification of worlds into nomically possible

problems in physics are undecidable. That is, there is no algorithm that can determine an answer to the problem in general (Richardson 1969, Cubitt, Perez-Garcia, and Wolf 2015a, Cubitt, Perez-Garcia, and Wolf 2015b, Cardona et al. 2021). For example, Terence Tao (2016) has developed a new approach to the Navier–Stokes global regularity problem – one of the famed Clay Millennium Prize problems. Within this program, Cardona et al. (2021) have found cases where determining the path of a particle in an ideal fluid is undecidable. Many other problems that naturally arise in physics have also been shown to be undecidable. See, for example, the following list of undecidable problems in physics from Cardona et al. (2021:1).

“several physical processes have been shown to exhibit such Turing completeness, from ray-tracing problems in three-dimensional (3D) optical systems Reif, Tygar, and Yoshida (1994) to neural networks (Siegelmann, Horne, and Giles 1997) or nonabelian topological quantum-field theories (Freedman 1998) For instance, the spectral gap problem (checking if the Hamiltonian of a quantum many-body system has a spectral gap) has recently been proved to be undecidable (Cubitt, Perez-Garcia, and Wolf 2022). Other undecidable problems in physics are the stability of an n -body system (Moore 1990), the problem of finding an Einstein metric for a fixed fourfold as observed by Wolfram (Wolfram 1985), or the reachability problem in potential well dynamics (Tao 2017).

When discussing so-called ‘Gödel phenomena’ in physics, it is important to clarify whether one is considering incompleteness or undecidability.

and nomically possible: we take a world to be nomically possible just in case it is a model of the theory.

I have presented a semantics of *cause* and *because* that, as far as the laws of nature are concerned, only requires a classification of worlds into nomically possible and nomically impossible. The proof-theoretic view of laws is compatible with this approach, since it determines such a classification. What we have seen, then, is that a proof-theoretic view of the laws is not required to model the truth conditions of *cause* and *because*. This is a welcome result, since it means that our analysis of the truth conditions of causal claims avoids gratuitous commitments concerning the strength of the mathematical theories that physicists may use.

6.4.2 The generality of nomic possibility

Our formalisation of nomic possibility is remarkably simple. It merely says, for each logically possible world, whether or not it is nomically possible. The formal simplicity of nomic possibility also ensures its generality: the simpler the notion, the easier it is for a wide variety of models to express it, and so the easier it is for models from various sciences to count as causal models in our sense.

One motivation for showing how various models generate a set of nomic possibilities comes from aspirations for the unity of science. One goal that unites all natural sciences is the search for the causal structure of the world. Sociologists, psychologists, biologists, chemists and physicists all aim to understand why things happen in the way they do: to utter true claims containing the words *cause* and *because*. If there is one thing, then, with the potential to unify science, it is causality.

To achieve this unification, the various models that scientists create should count as causal models. Given the simplicity of the notion of nomic possibility, many models do determine a set of nomic possibilities. Here are some examples of models that do, from mathematics, physics, computation, statistics and artificial intelligence.

1. *Dynamical systems* (Poincaré 1890b, 1899). Given a dynamical system over a set X , a world is a set of elements of X indexed by time. A world is nomically possible just in case it is a solution to the system's evolution rule.¹⁵
2. *Models of general relativity* (Einstein 1916). As we discuss in section 6.4.3, we can take a world to be a four-dimensional Lorentzian manifold. A world

¹⁵Formally, a *continuous time dynamical system* is a pair (X, φ) where X is a set and φ a family of maps $\varphi_t : X \rightarrow X$ with $t \in \mathbb{R}$, such that φ_0 is the identity map and $\varphi_{t+s} = \varphi_t \circ \varphi_s$ for all $t, s \in \mathbb{R}$. A path is a set $\{x_t : x \in X, t \in \mathbb{R}\}$. A *solution* to a dynamical system is a path p such that $x_t = \varphi_t(x_0)$ for all $x_t \in p$. Here we have defined a *flow* – analogous definitions can be given for discrete dynamical systems and semiflows. For a quick overview of dynamical systems see Barreira and Valls (2012: chapter 3).

is nomically possible according to general relativity just in case it satisfies the Einstein field equations.

3. *Turing machines* (Turing 1936). We take a world to be a sequence of states of a Turing machine; each state at a moment in time specifies the state of the tape, the position of the head, and the machine's internal state. A world is nomically possible just in case states evolve according to the machine's state transition function.

By the Church-Turing thesis, every model of computation whatsoever can be represented as a Turing machine. Since each Turing machine determines a set of nomically possible worlds, every model of computation does too.

4. *Bayesian networks and structural equation models* (Pearl 1988, 2000, Spirtes, Glymour, and Scheines 1993). In section 6.5 below we show that from each Bayesian network and structural equation model we can determine a set of nomically possible worlds.

From the perspectives of linguistics and cognitive science, representing relativity is not a major concern. After all, their task is to represent human reasoning, and humans generally go about the world believing in absolute simultaneity – that everyone has the same ‘now’. But apart from the unity of science, there is another reason why causal modellers should care about models of physics in particular. An overarching question in philosophy is to understand the relationship between thought and the world. In terms of the present project, we would like to understand what makes our causal claims true or false. The present approach provides one answer to this question: it is what our model represents, which includes a notion of nomic possibility. Physics currently provides our best understanding of the laws of nature, so we would like to be able to relate the models we have assumed with models in physics, such as models of general relativity.

6.4.3 Representing models of physics

A top down model construction. Our model construction in section 6.1.1 built a linear order of time into the construction of worlds. There we took what we may call a ‘bottom up’ approach, starting with the smallest parts of the model, states, and building up from there. If we wish to take worlds as primitive – a common practice in possible worlds semantics – we may take a ‘worlds first’ or ‘top down’ approach, beginning with worlds and assigning structure to them as needed. On this approach, a model has the form

$$(W, P, \mathcal{T}, S, \leq, \mathcal{M}, \mathcal{A}, |\cdot|)$$

where W is a set, \mathcal{T} is a function assigning to each world a linear order, and M is a function assigning to each world and element of its linear order a maximal

element from (S, \leq) . (As above, (S, \leq) is a state space, \mathcal{A} is a relation between states and sentences, P a set of worlds and $|\cdot|$ a function from worlds to sentences.) Intuitively, W represents the set of possible worlds, T assigns temporal structure to worlds, and \mathcal{M} assigns mereological structure to worlds at a moment in time.

A top down construction is useful since it allows to determine nomic possibility with respect to worlds independently of their temporal and mereological structure. By separating nomic possibility from time and mereology, the top down construction allows us to treat these separate components differently. For example, it is often observed that the laws of nature are time symmetric while our experience of time is asymmetric. On this view, one might think the laws of nature are objective while the direction of time is subjective, a product of our cognition (perhaps in association with the second law of thermodynamics). With the top down construction, we can easily separate our model into objective and subjective components along these lines; say, classifying W and P as the mind-independent part of the model and $(S, \leq, \mathcal{T}, \mathcal{M}, \mathcal{A}, |\cdot|)$ as the mind-dependent part of the model. Different views carve up the model into mind-dependent and mind-independent in different ways. These carvings can be expressed on the top down construction but not on the bottom up construction.

While the bottom up construction is simpler to state, it builds a linear direction of time into the construction of worlds. This creates a difficulty when it comes to generating a set of nomic possibilities from relativity. In contrast, the top down construction helps us appreciate the relationship between the present modelling framework and models of relativity. Einstein would not be happy with a bottom up approach, where worlds are constructed out of simultaneous states.

In relativity, we define what it means for a world to be nomically possible independently of simultaneity: a world is nomically possible just in case it satisfies the laws of general relativity (the Einstein field equations, where a world is a four-dimensional Lorenzian manifold). Simultaneity does not come from the laws of nature themselves but from an observer. On a bottom up approach, however, worlds and simultaneity come packaged together as one. To represent models of general relativity on a bottom up approach, we would first need to separate worlds from their temporal structure before we can classify which worlds are nomically possible. Representing models of relativity in the present framework is far simpler on a top down approach.

Now that we understand the present answer to the modelling question, let us discuss why we have chosen this model over alternatives. In chapter 3 we discussed the state space and aboutness relations, showing how they allow us to represent how we imagine a sudden change to imagine a sentence true. We also saw a plausible way to analyse aboutness in terms of the state space and nomic possibility (whereby an atomic sentence A is about state s just in case s minimally settles A , and aboutness for logically complex sentences is determined compositionally from their logical form). This reduction would allow us to simplify our model to only four components: $(S, \leq, P, |\cdot|)$. As we have already discussed the

state space and aboutness at length, we will not further consider them here. Nor will we discuss the interpretation function, since it is not unique to our model but part of every truth-conditional semantics whatsoever. That leaves nomic possibility. It will be our focus for the remainder of this chapter.

6.4.4 Comparison with alternatives to nomic possibility

In this section we compare this representation of the laws of nature with two others: propositional logic, on the one hand, and Bayesian networks and structural causal models, on the other.

To fix an example, consider how we should represent the behaviour of a switch and light. A first thought might be to use propositional logic. We could take the law governing the system's behaviour to be $L \leftrightarrow S$, where S represents that the switch is down, L that the light is on, and \leftrightarrow is material equivalence (i.e. $L \leftrightarrow S$ true just in case S and L are both true or both false). An immediate problem is that this formula does not capture causal asymmetry, since equivalence is symmetric: $L \leftrightarrow S$ is equivalent to $S \leftrightarrow L$. There is, however, an even more fundamental problem with this representation, one that comes from examining its relationship with time. Laws are supposed to be invariant; in particular, they should hold at all times. What the law $L \leftrightarrow S$ actually says, then, is that for every time t , the light is on at t if and only if the switch is down at t . Now, this is simply not true. There is a time just after one flicks the switch when the light has yet to change. However slight the delay is, it nonetheless exists. If the delay is so short that we consider it negligible, we may replace the example with one with a longer delay (say, with an especially slow bulb).

The existence of the delay means we have to reevaluate in what sense we took $L \leftrightarrow S$ to correctly describe the relationship between the switch and light. That formula is true most of the time, but not always. To regard $L \leftrightarrow S$ as a true description of the behaviour of the switch and light we must turn a blind eye to times after we flick the switch but before the light changes. $L \leftrightarrow S$ is approximately true. The problem is that we are not looking for an approximately true model – we are looking for a true model.¹⁶

One may reply that when we take $L \leftrightarrow S$ to be a true description of the relationship between the switch and light, we have in mind a distinction between the unchanging states and the changing states, and are implicitly restricting attention

¹⁶One reason why $L \leftrightarrow S$ may have seemed like a true description of the switch and light system is that logical formulas are typically used to represent unchanging things. This reflects the origin of modern logic as a means to describe mathematical objects, which on the Platonic conception are taken to exist outside time (a notable alternative view is Brouwer's 1948 idea of the creating subject, though this is not how most mathematicians think of the nature of mathematical objects). In light of this tradition, when we see a logical formula we are conditioned, I believe, to implicitly assume that the things it describes are static. This, of course, is a major obstacle when giving a formal analysis of causality, where change is of central concern.

to the unchanging states. In this case the unchanging states are assumed to be the states where L and S are both true or both false. But this classification is not quite correct. It is perfectly possible for a so-called ‘unchanging’ state to change; say, for someone to flick the switch. Since every state is capable of changing, a restriction to the states that cannot change would leave no states at all.

One way to fix this situation is to read the formula $L \leftrightarrow S$ asymmetrically. There are many conceivable ways to implement this asymmetry. A popular approach today is to model the behaviour of the switch and light using functional dependence. This is assumed by Bayesian networks (Pearl 1988) and structural equation models (Wright 1921, Haavelmo 1943, Koopmans 1950) and structural causal models (Pearl 2000). On this approach we represent the switch and light as binary variables, and require that the value of the light is a function of the value of switch: $L = f(S)$.¹⁷ This gives us the desired asymmetry: while $L \leftrightarrow S$ is equivalent to $S \leftrightarrow L$, a function taking values of the switch and returning values of the light is not the same as a function taking values of the light and returning values of the switch: $L = f(S)$ is not equivalent to $S = f(L)$.

To illustrate, we are quite comfortable representing the switch and light with the structural causal model in Figure 6.21, where the equation $L = S$ is read from left-to-right, saying that the state of the switch is given independently, and the state of the light is a function of the state of the switch.

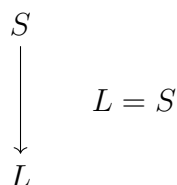


Figure 6.21: A structural causal model of the switch and light.

This solves the symmetry problem, but does it avoid the false prediction of the formula $L \leftrightarrow S$, that for every time t the light is on at t if and only if the switch is up t ? A great deal of work has gone into investigating the testable predictions of structural causal models, and Bayesian networks more generally – a structural causal model being a special case of Bayesian networks, where the probabilities are 0 or 1, and we ignore the probabilities of the exogenous variables. Figure 6.22 gives the structural causal model as a Bayesian network. Each Bayesian network generates a probability distribution over assignments of values to variables. This Bayesian network generates the distribution in Table 6.1.

Now let us imagine that on one hundred occasions we peeked into the room to record the status of the switch and light. On two occasions, let’s say, we found

¹⁷Strictly speaking, $L = f(S)$ is shorthand for the claim that the model contains a function $f : \mathcal{R}(S) \rightarrow \mathcal{R}(L)$ from the range of values of the switch to the range of values of the light. In this case we have $f : \{\text{up}, \text{down}\} \rightarrow \{\text{on}, \text{off}\}$ where $f(\text{up}) = \text{off}$ and $f(\text{down}) = \text{on}$.

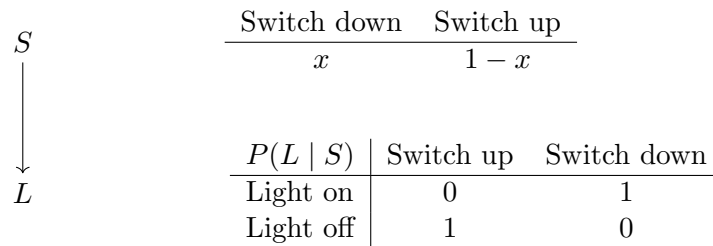


Figure 6.22: A Bayesian network representing the switch and light.

		Probability
switch down	light off	0
switch down	light on	x
switch up	light off	$1 - x$
switch up	light on	0

Table 6.1

that the switch was down but the light was still off. We happened to peek in just after someone had flicked the switch. Our data given in Table 6.2.

		Observations
switch down	light off	2
switch down	light on	48
switch up	light off	0
switch up	light on	50

Table 6.2

This database is incompatible with the Bayesian network in Figure 6.22. It faces the same problem as our formula $L \leftrightarrow S$. It is only a correct model of the relationship between the switch and light when we ignore some states – when we erase some of our data.

One could respond that our data is a mere approximation of the underlying distribution that generates the data. Like almost all data sets, it contains some measurement error. One must ask: when we looked into the room just after someone had flicked the switch, but the light had yet to change, were we performing a measurement error? Our task was to record the state of the switch and light a number of times. Looking at the switch and light every now and then seems to be a perfect way to do that. (If one thinks that the modellers should have observed the switch and light for a longer period of time, the question then becomes how they should decide which states of the switch and the light during that period of observation to record. The obvious answer is to record all that they see.)

A second issue with this response is that, with more data and better data collection methods, we expect errors to vanish in the long run. But given that it

takes some time for the light to change after the switch is flicked, no matter how much data we collect we will always observe states where the switch has been flicked but the light has yet to change. If this is an error, it is one that never goes away.

These remarks make clear that, when we use propositional logic, or a Bayesian network or structural causal model to model some system (such as the behaviour of the switch and light), we ignore some states. We only count those where the system is in equilibrium.

With these points I do not wish to imply that structural causal models and Bayesian networks are based on false assumptions, or cannot be used to represent the relationship between the switch and light. I merely wish to call attention to some implicit assumptions we must make when we apply these frameworks; assumptions about what we mean when we say that a structural causal model or Bayesian network is a correct model of a scenario.

6.5 A dynamic interpretation of structural causal models

Structural causal models (Pearl 2000) are a popular framework in which to model causal relations. The goal of this section is to show that there is a natural way to represent structural causal models in the present framework. We show that every structural causal model generates a partial model of the nomic possibilities. I call this a *dynamic interpretation* of structural causal models since time will play a starring role (though it is not the only essential component: nomic possibility will be central too).

We begin with a definition of structural causal models, from Pearl (2009:203).

6.5.1. DEFINITION. A *structural causal model* is tuple (U, V, R, F) where

- U and V are disjoint sets of variables, called *exogenous* and *endogenous*, respectively.
- R assigns to each variable in $U \cup V$ a set of values. We extend R to sets of variables by letting $R(\{X, Y, \dots\}) = R(X) \times R(Y) \times \dots$ for any set of variables $\{X, Y, \dots\}$.¹⁸
- F assigns to each endogenous variable $X \in V$ a function $f_X : R(PA_X) \rightarrow R(X)$ where $PA_X \subseteq U \cup V \setminus \{X\}$.

To illustrate with our running example of the switch and light, Figure 6.23 represents their behaviour as a structural causal model, together with its representation as a graph. $S = 0$ denotes that the switch is up, $S = 1$ that it is down, $L = 0$ that the light is off, and $L = 1$ that it is on.

¹⁸Strictly speaking, this definition requires assuming an arbitrary linear order over the vari-

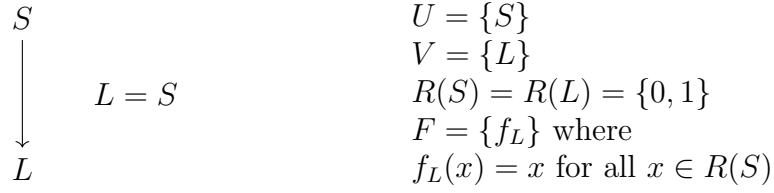


Figure 6.23: The switch and light as a structural causal model.

We can give structural causal models a dynamic interpretation, translating them into partial models. Let us first give the dynamic interpretation, and then show why it is useful.

6.5.1 The dynamic interpretation

The dynamic interpretation is quite simple: we interpret the functions as descriptions of how the values of the variables change in time in response to the values of their parents. We view the functions' inputs as preceding their output by one step in time. In other words, we interpret the functions as the transition functions of a discrete dynamical system.

Given a structural causal model $M = (U, V, F, R)$, let us define that a *state* of M is an assignment of values to the variables; that is, a function $s : U \cup V \rightarrow R(U \cup V)$. A *path* of M is a sequence of states (s_0, s_1, \dots) where each s_t is a state of M . We can take time to be infinite into the past or future as we please. Here we will take paths to have a starting point and continue infinitely into the future, though nothing hinges on this point. So our paths have the form $(s_t)_{t \in \mathbb{N}}$, where $\mathbb{N} = \{0, 1, \dots\}$ is the set of natural numbers.

A path of M is *nomically possible* just in case the states transition according to the structural equations; that is, just in case we have

$$s_{t+1}(X) = f_X(s_t(PA_X))$$

for all $t \in \mathbb{N}$ and $X \in V$. Note the use of time in this equation. The value of X 's parents at a given state determine the value of X at the next state, according to the structural equation for X . Note also that the equation ranges only over the endogenous variables. The result is that the exogenous variables can change freely from state to state. This represents the fact that the exogenous variables are determined by factors outside the scope of the model.¹⁹

ables, since cross-products are ordered while sets are not; $\{X, Y\} = \{Y, X\}$ but in general $R(X) \times R(Y) \neq R(Y) \times R(X)$.

¹⁹For another interpretation of structural equations using dynamics, see Schulz (2011). Our dynamic interpretation differs from Schulz's in some small ways. Schulz allows variables to have undetermined values, restricts attention to variables with finitely many parents, and does not allow the background variables (what we call the exogenous variables) to change value.

Putting all this together, we define the dynamic interpretation of structural causal models as follows.

6.5.2. DEFINITION (Dynamic interpretation). For any structural causal model $M = (U, V, R, F)$, define the *paths* of M to be

$$\text{paths}(M) = \{(s_t)_{t \in \mathbb{N}} \mid s_t : U \cup V \rightarrow R(U \cup V) \text{ for all } t \in \mathbb{N}\}.$$

The *dynamic interpretation* of M is the pair $\mathcal{DI}(M) = (P^+, P^-)$ given by

$$\begin{aligned} P^+ &= \{(s_t)_{t \in \mathbb{N}} \in \text{paths}(M) \mid s_{t+1}(X) = f_X(s_t(PA_X)) \text{ for all } t \in \mathbb{N} \text{ and } X \in V\} \\ P^- &= \text{paths}(M) \setminus P^+ \end{aligned}$$

For example, take the structural causal model of the switch and the light. It turns out that its dynamic interpretation is exactly the model we discussed in section 6.2.3, where the paths of the model are the infinite paths through Figure 6.15, repeated in Figure 6.24 below, and a path is nomically possible just in case it always follows the arrows.

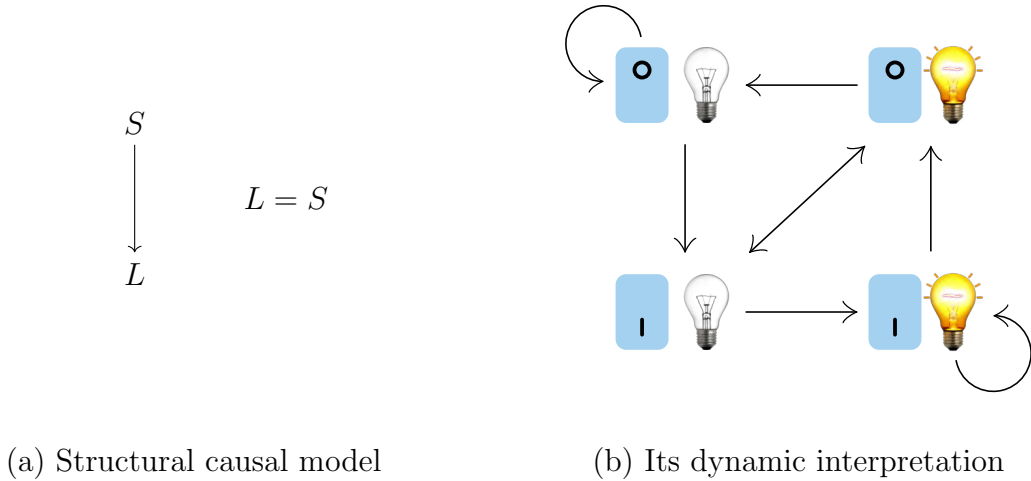


Figure 6.24

The figure illustrates a general fact about the dynamic interpretation of models with finitely many variables: the states that satisfy the structural equations are those with loops, i.e. those that can lawfully succeed themselves. This follows from the fact that a state s_t satisfies the structural equations just in case $s_t(X) = f_X(s_t(PA_X))$ for every variable X , which by our definition of the dynamic interpretation is exactly what we need for a transition from s_t to itself to be nomically possible.

Schulz also uses the dynamics to give a semantics of counterfactuals, whereas our semantics of counterfactuals in chapter 3 does not use the dynamic interpretation of structural causal models.

Here are some sample calculations to show how we arrived at this interpretation. Where $s_t = (0, 0)$ and $s_{t+1} = (0, 1)$, the derivation on the left shows that it is not nomically possible for the light to spontaneously turn on. And where $s'_t = (1, 0)$ and $s'_{t+1} = (1, 1)$, the derivation on the right shows that it is nomically possible for the light to turn after the switch has been flicked.

$$\begin{array}{llll}
 s_{t+1}(L) & = & f_L(s_t(PA_L)) & s'_{t+1}(L) & = & f_L(s'_t(PA_L)) \\
 \text{iff} & 1 & = & f_L(s_t(S)) & \text{iff} & 1 & = & f_L(s'_t(S)) \\
 \text{iff} & 1 & = & f_L(0) & \text{iff} & 1 & = & f_L(1) \\
 \text{iff} & 1 & = & 0 & \text{iff} & 1 & = & 1
 \end{array}$$

Let us explore some further examples of dynamic interpretations. Figure 6.25 depicts a chain $X \rightarrow Y \rightarrow Z$ and its dynamic interpretation. Notice the sequential behaviour: it takes two steps in time for a change in the value of X to change the value of Z . Figure 6.26 shows that the dynamic translation also applies to structural causal models with cycles (also known as non-recursive models). Given binary variables X and Y and structural equations $X = Y$ and $Y = X$, the dynamic interpretation predicts that the nomically possible paths are those where X and Y are always true, always false, and those that oscillate between true and false:

$$P^+ = \left\{ (\overline{xy}, \overline{xy}, \overline{xy}, \dots), (\overline{xy}, xy, \overline{xy}, \dots), \right. \\
 \left. (x\overline{y}, \overline{xy}, x\overline{y}, \dots), (xy, xy, xy, \dots) \right\}$$

Finally, Figure 6.27 depicts the dynamic translation of an AND-gate. The complexity of the dynamic translation illustrates how structural causal models express information in a remarkably compact way. This illustrates a major benefit of using structural causal models (SCMs) to express dynamic information: they provide a simple representation of a complex concept (functional dependence). In contrast, the dynamic interpretation provides a complex representation of a simple concept (nomic possibility). Thus presenting an SCM is more appropriate when one wishes to communicate a large volume of high-level information simply, while presenting its dynamic interpretation is more appropriate when one is in a philosophical mood, wishing to understand what features of our experience the SCM really corresponds to. As Quine writes, “It is one of the consolations of philosophy that the benefit of showing how to dispense with a concept does not hinge on dispensing with it” (Quine 1960:189). Each representation has its place.

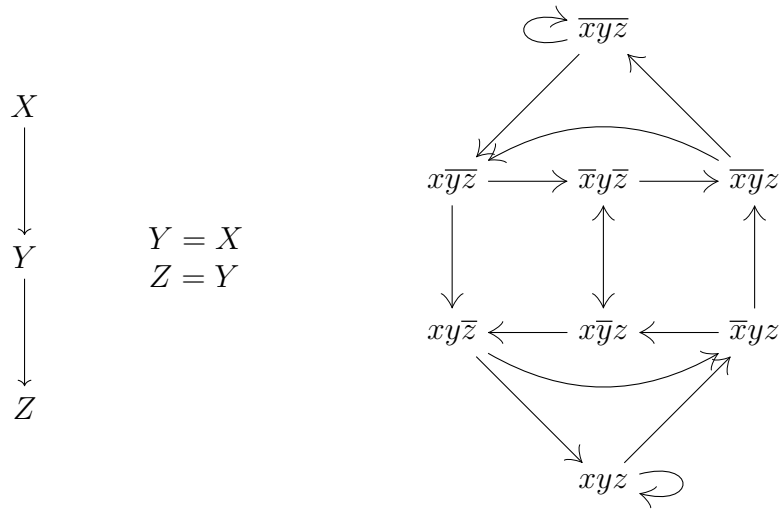


Figure 6.25: A chain and its dynamic interpretation, where X , Y and Z are binary variables. The point $x\bar{y}z$, for example, represents the state where $X = 1$, $Y = 0$ and $Z = 1$.



Figure 6.26: The dynamic interpretation of a cyclic structural causal model.

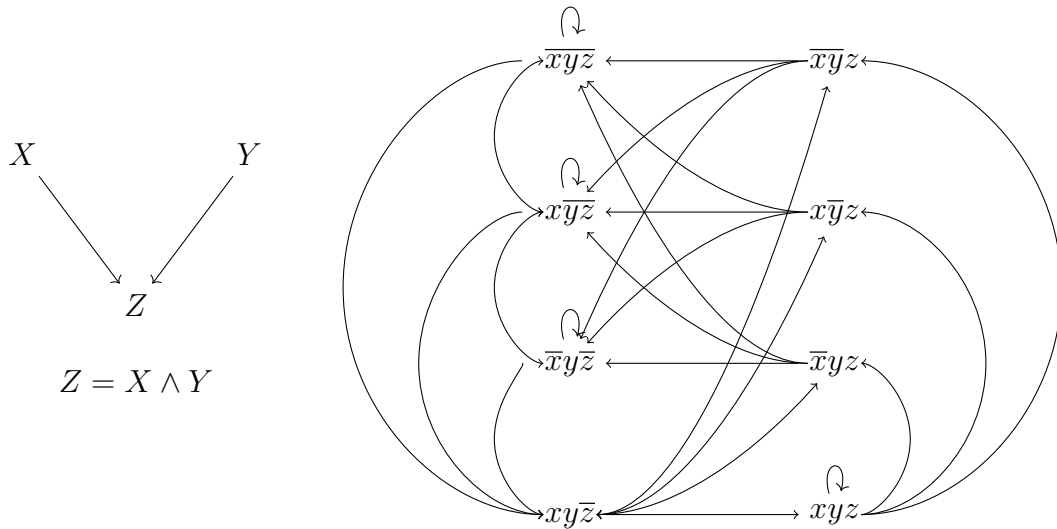


Figure 6.27: An AND-gate as a structural causal model (left), and its dynamic interpretation (right).

6.5.2 The need for the dynamic interpretation

An overarching question facing those who use structural causal models is to say what it means for a structural causal model to be true.²⁰ What, in the world, is a structural causal model? To illustrate, suppose someone comes along who insists that the correct model of the switch and light has the reverse dependence: they believe that the value of the switch should be represented a function of the value of the light, not the other way around. What would we say to them to convince them of their mistake? In other words, how do we come to decide that model (a) below is a correct model of the switch and light and that model (b) is incorrect? What does the correctness of model (a) correspond to in our experience?



Figure 6.28: How do we decide that (a) is correct and (b) incorrect?

The dynamic interpretation will provide an answer to this question: a structural causal model is true just in case its dynamic interpretation is true. And where P is the true set of nomically possible worlds and $\mathcal{DI}(M)$ the dynamic interpretation of an SCM M , Definition 6.2.14 tells us that $\mathcal{DI}(M)$ is true at P just in case every path that is possible according to $\mathcal{DI}(M)$ is the abstraction of a world in P , and no path that is impossible according to $\mathcal{DI}(M)$ is the abstraction of a world in P . As we saw in section 6.4.2, this provides a remarkably general answer to the question of what it means for an SCM to be true, one that builds a bridge between the SCM modelling framework and others in a wide variety of domains, from mathematics (dynamical systems) and physics (models of general relativity) to computer science (models of computation).

Thus the model (a) is correct since it correctly describes the nomic possibilities and impossibilities: how the switch and light can change through time. Model (b) is incorrect since it does not. As we will see, the dynamic interpretation of (b) incorrectly implies that the light can spontaneously turn on, without there

²⁰There is a separate question, concerning what it means for a structural causal model to be *apt*. This question arises when one wishes to use structural causal models to predict the truth conditions of counterfactual and causal claims. Since the present approach do not use structural causal models to do so, the question of aptness does not arise here. For discussions of aptness, see Blanchard and Schaffer (2017:181–83), Hall (2007), Halpern and Hitchcock (2010), Woodward (2016b), Menzies (2017), and McDonald (2022).

first being a change in the switch, which allows us to conclude that model (b) is incorrect.

Let us review some alternative answers to the question of what it means for a causal model to be true. One approach is to point to manipulation:

one says “*A* causes *B*” in cases where one could produce an event or state of the *A* sort as a means to producing one of the *B* sort.

(Gasking 1955:485)

The paradigmatic assertion in causal relationships is that manipulation of a cause will result in the manipulation of an effect. ... Causation implies that by varying one factor I can make another vary.

(Cook and Campbell 1979:36)

Rubin and Holland summarise the view with their slogan, “No Causation without Manipulation” (Holland 1986).²¹ This approach is often coupled with human agency (Von Wright 1971, Price 1991, 1992). For example:

... an event *A* is a cause of a distinct event *B* just in case bringing about the occurrence of *A* would be an effective means by which a free agent could bring about the occurrence of *B*.

(Menzies and Price 1993:187)

There are well-known challenges to this view (see Woodward 2016a for discussion). It is particularly suited to causation at the human scale, but does not generalise well beyond that (see e.g. Pearl 2000:361). For example, the big bang caused stars to form, but the claim that bringing about the big bang is ‘effective means’ by which a free agent could make stars form is only true in a loose, highly attenuated metaphorical sense. But the claim that the big bang caused stars to form is strictly true, not merely metaphorically true. If one replies that our judgement here is a result of analogy, of thinking what would happen were there some omnipotent agent with an effective means to make the big bang happen, the question then becomes to pinpoint what features of the scenario we use to make this judgement. The framework proposed in this thesis provides an answer to this question: it provides a semantics of *cause* which tells us what it means for “The big bang caused stars to form” to be true in terms of the models we have developed.

Nonetheless, the present framework plausibly also has the potential to account for those cases where it is appropriate to describe causal relations in terms of manipulation. For example, it is plausible to say that flicking the switch is an effective means of turning the light on because flicking the switch is sufficient to produce the light to turn on, and that the light being on is not an effective

²¹For an overview of manipulation approaches to causation see Woodward (2016a).

means of the switch being up because the light being on is not sufficient to produce the switch to be up. And as we saw in chapters 3 and 5, we can analyse sufficiency and production in the present framework. This analysis of ‘effective means’ would of course have to be justified in greater detail, but it suggests that manipulation-based accounts of structural causal models can be expressed in the present framework, while the present framework can also go beyond manipulation-based accounts in cases where they do not apply, such as at the cosmological scale.

Moving away from agency, an alternative idea is that causal dependencies represent mechanisms:

Each parent–child relationship in the network represents a stable and autonomous physical mechanism.

(Pearl 2000:22)

Each equation represents a distinct mechanism (or law) in the world, one that may be modified (by external actions) without altering the others.

(Halpern and Pearl 2005:847)

This is a helpful way to illustrate the framework of structural causal models. But without a precise understanding of what a mechanism is, this does not tell us whether a structural causal model is correct or not, for it does not tell us what it means in general for an equation to correctly represent a mechanism or law in the world.²² The idea of a mechanism is quite a high-level concept. If someone claimed that model (b) is a correct description of the mechanism of the switch and light, what more basic facts would we point to, to convince them of their mistake? We might reply that they simply have the ‘wrong’ concept of mechanism, that they misunderstand the concept. This response would likely alienate them, and would certainly not teach them where they have gone wrong.

In contrast, the dynamic interpretation gives us a simple way to respond. For the dynamic translations of models (a) and (b), respectively, are given by the topmost diagrams of Figure 6.17. As we can see, these make different predictions about which worlds are nomically possible. Model (a) predicts that it is nomically possible for the switch to change without there first being a change in the light; model (b) predicts this to be impossible. To one who claims that model (b) is correct, we can simply have them flick the switch and see the light then turn on. Since actuality implies possibility, this shows that model (b) wrongly predicts something to be impossible which is in fact possible. This is a proof from direct experience that model (b) is incorrect. It is a proof we can all agree on.

The present framework also relies on simple concepts. To illustrate, let us describe what someone who observes the switch and light experiences; a toddler,

²²For an influential discussion of the concept of mechanism, see Machamer, Darden, and Craver (2000).

say, in the same room as the switch and light. We can imagine they are too short to reach the switch, or do not have the dexterity to manipulate the light switch – their fine-motor skills are not yet developed enough. They have no way to effectively intervene on the system. Moreover, they have no idea what mechanism underlies the switch and light – they do not have the concept of electricity, they do not know that there are wires hidden in the walls, and so on. Lastly, they do not yet have the sophisticated linguistic abilities required to interpret counterfactuals.²³ However, the toddler can observe whether the switch is up or down and whether the light is on or off. And of course, the toddler experiences the world in time (experience outside time, as any reader of Kant will tell you, is impossible).

Now imagine it is evening: the room is dim. The toddler notices that the switch is up and the light is off. An adult walks into the room. The child notices them flick the switch down; for a brief moment the room is still dark. Then the light turns on. Sometime later it is bedtime. The child sees the adult flick the switch up. The light turns off. Figure 6.29 illustrates the toddler’s experience of the switch and light.



Figure 6.29: Some of the toddler’s experience.

After some time the toddler sees many sequences of states of the switch and light. Suppose they jump to the conclusion that the sequences they have seen are the only possible sequences of the switch and the light. The toddler then has an opinions about what sequences of states are possible and which are impossible. Thus the toddler has a (partial) model of the nomic possibilities.

Every definition of truth naturally furnishes a definition of entailment. Since the dynamic interpretation will provide an analysis of what it means for a SCM to be true (with respect to the set of nomically possible worlds), it will also provide an analysis of what it means for one SCM to entail another. Given two SCMs M and M' , we may say that M entails M' just in case the dynamic interpretation of M entails that of M' , with entailment defined as in Definition 6.2.12.²⁴

²³The acquisition of conditional constructions happens comparatively late in linguistic development; see Reilly (1982), Bowerman (1986), Nyhout and Ganea (2019), and Tulling and Cournane (2019).

²⁴An exciting avenue for future work is to compare this notion of entailment with recent work concerning abstraction of structural causal models (see note 5).

6.5.3 Previous hints at the dynamic interpretation

Among those who have developed the structural causal modelling framework, there is a persistent tendency to read the functional dependencies as expressing dependence in time. Here are four examples. Firstly, Pearl writes that

The choice of PA_i (connoting *parents*) is not arbitrary, but expresses the modeller’s understanding of which variables Nature must consult **before** deciding the value of V_i .

(Pearl 2009:203, note 3, my emphasis).

Secondly, in Pearl, Glymour, and Jewell’s textbook *Causal Inference in Statistics*, when the authors discuss colliders (graphs of the form $X \rightarrow Z \leftarrow Y$) they write that the independence of X and Y

reflects our understanding of how causation operates in time; events that are independent in the present do not become dependent merely because they may have common effects in the future.

(Pearl, Glymour, and Jewell 2016:41).

Thirdly, after Pearl (2009:37) introduces his three-step procedure for evaluating counterfactuals $[X = x]Y = y$ in structural causal models, he writes:

In temporal metaphors, this three-step procedure can be interpreted as follows. Step 1 explains the past (U) in light of the current evidence e ; step 2 bends the course of history (minimally) to comply with the hypothetical condition $X = x$; finally, step 3 predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

Finally, consider the choice of the term ‘Markov’. Traditionally, a discrete stochastic process is called Markov just in case it is memoryless: every event is probabilistically independent of the previous states conditional on the state of its immediate successor. Pearl and Paz (1985) later applied the term to Bayesian networks. A Bayesian network satisfies the Markov condition just in case every variable is independent of its non-descendants conditional on its parents (Spirtes, Glymour, and Scheines 1993:11). The choice of the term Markov is telling. It suggests that a variable’s non-descendants represent its past, with its parents representing its immediate past.

It is natural to wonder about the status of these many temporal metaphors. Why do the architects of these frameworks consistently find it useful to speak of them in temporal terms? The dynamic interpretation offers a way to make sense of these metaphors: they are not metaphors at all, but reflect what it means for a structural causal model to be true.

6.5.4 The exogenous/endogenous distinction

An important difference between structural causal models and their dynamic interpretations is that structural causal models explicitly distinguish the exogenous and endogenous variables, while their dynamic translations do not. This is a plausible result, since the exogenous/endogenous distinction arguably reflects the interests of the modeller rather than a distinction in reality.

While the dynamic interpretation of a structural causal model does not officially distinguish the exogenous and endogenous variables. Given the dynamic interpretation of a structural causal model M , it is easy to recover which variables are exogenous and which are endogenous in M . Exogeneity resurfaces as randomness: the exogenous variables are those that can change freely: variable X is exogenous in a structural causal model M just in case in the dynamic interpretation of M , for every state s_t and possible value x of X , s_t has a successor where X has value x . This reflects the fact that the exogenous variables are those whose behaviour is beyond the scope of the model. Accordingly, the dynamic interpretation of a model does not impose any constraints on how the exogenous variables they may change, besides predicting that each of their values is possible (an uncontroversial commitment given that the range of each variable represents the possible values it may take).

Since the behaviour of the exogenous variables are independent of the structural equations, when we present the dynamic interpretation of a structural causal model it can be convenient to ignore changes of the exogenous variables. That is, we can present the interpretation while adding the requirement that for every nomically possible path $(s_t)_{t \in \mathbb{N}}$ we have $s_{t+1}(U) = s_t(U)$ for any exogenous variable U . Adding this restriction allows us to simplify the representation of dynamic interpretations, as shown in Figures 6.30 and 6.31.

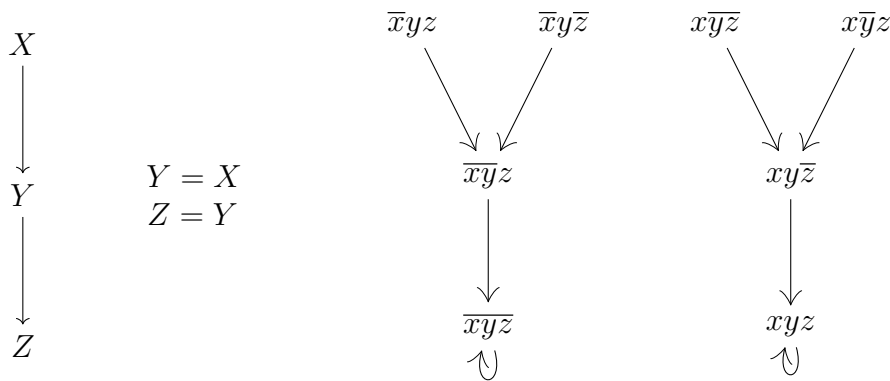


Figure 6.30: The dynamic interpretation of a chain, with exogenous variables fixed.

Formally, let M be a structural causal model and $\mathcal{DI}(M) = (P^+, P^-)$ the

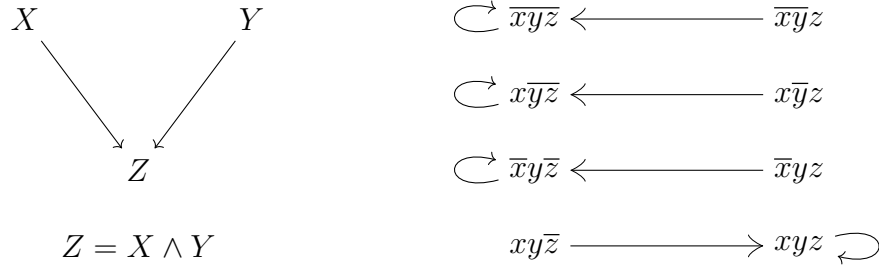


Figure 6.31: An AND gate as a structural causal model (left), and its dynamic interpretation with exogenous variables fixed (right).

dynamic interpretation of M . Then the dynamic interpretation of M with the exogenous variables fixed, denoted $\mathcal{DI}_{fixExo}(M)$, is the pair $(P_{fixExo}^+, P_{fixExo}^-)$ given by

$$\begin{aligned} P_{fixExo}^+ &= \{(s_t)_{t \in \mathbb{N}} \in P^+ \mid s_{t+1}(X) = s_t(X) \text{ for all } t \in \mathbb{N} \text{ and } X \in U\} \\ P_{fixExo}^- &= paths(M) \setminus P_{fixExo}^+ \end{aligned}$$

6.5.5 Transfinite dynamic interpretations

An interesting question for the dynamic interpretation is what to do when a variable has an infinite chain of parents; that is, when there is a variable Y and an infinite chain of variables $(X_n)_{n \in \mathbb{N}}$ such that each X_n is a parent of both X_{n+1} and Y . Figure 6.32 gives an example of structural causal model with such a chain, where each X_n and Y are binary variables.

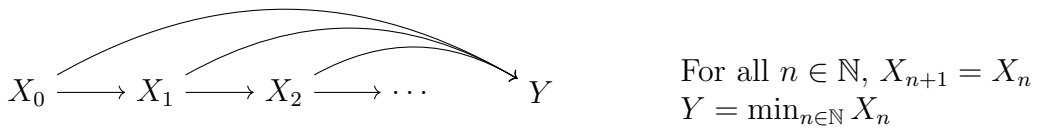


Figure 6.32

There are two ways to give a dynamic interpretation of such a model. The first is to assume that time is not transfinite and stick with Definition 6.5.2. (Formally, time is transfinite just in case there is a nonzero time t such that for any time $t' < t$, there is a time t'' with $t' < t'' < t$.) One might have a philosophical objection to transfinite time and therefore prefer this route. In this case, in the dynamic interpretation of the model in Figure 6.32 it is not possible for Y to change value. This means that, if we do not allow for transfinite paths, the model above has the same dynamical interpretation as one we replace the equation for Y with $Y = 0$.

Alternatively, we can allow time to be transfinite. This would allow Y to change value in the dynamic interpretation of Figure 6.32, after an infinite amount of time has passed. It would therefore allow us to distinguish the dynamic interpretations of the model above and the model where Y 's equation is replaced with $Y = 0$.

To construct a dynamic interpretation where time can be transfinite, we take a path to be a sequence of variable assignments indexed by ordinals. For time points at infinity (in set theory speak: at limit ordinals), the value of each variable is determined by the limit of the previous values of its parents. For time points not at infinity (that is, at successor ordinals) things are as usual: the value of each variable is determined by the values of its parents at the previous state.

Formally, we define that a path (s_0, s_1, \dots) is nomically possible just in case for any endogenous variable $X \in V$ and ordinal α ,

$$s_{\alpha+1}(X) = f_X(s_\alpha(PA_X))$$

and for any limit ordinal λ , the limit $\lim_{\alpha < \lambda} s_\alpha(PA_X)$ exists and

$$s_\lambda(X) = f_X\left(\lim_{\alpha < \lambda} s_\alpha(PA_X)\right)$$

where $\lim_{\alpha < \lambda} s_\alpha(PA_X) = \{ \lim_{\alpha < \lambda} s_\alpha(Y) : Y \in PA_X \}$.

For example, the limit of $(1, 0, 0, 0, \dots)$, $(1, 1, 0, 0, \dots)$, $(1, 1, 1, 0, \dots)$, \dots is $(1, 1, 1, 1, \dots)$, the sequence with all 1s. Applying the function for Y to this sequence returns $f_Y(1, 1, 1, 1, \dots) = 1$. So where s_0 assigns 1 to X_0 and 0 all other variables, and X_0 stays at 1 throughout the sequence, after an infinite amount of time Y changes from 0 to 1: $s_n(Y) = 0$ for every natural number n , but $s_\omega(Y) = 1$, as desired.²⁵ Thus there is a time when the structural equations in Figure 6.32 are all true.

In what follows we will assume that time can be transfinite and adopt the treatment of the dynamic interpretation just stated. The benefit of doing so is an increase in generality: we can provide a dynamic interpretation of graphs with infinite chains of parents. If one has philosophical objections to transfinite time, one can accordingly restrict attention to the dynamic interpretation of structural causal models without infinite chains of parents.

6.5.6 Recursive models and eventual truth

The dynamic interpretations of recursive structural causal models are especially well-behaved. Following Halpern and Pearl (2005:849), we call a structural causal model *recursive* just in case there is a strict total order \prec over the variables such

²⁵Note that the alternative scopal configuration $s_\lambda(X) = \lim_{\alpha < \lambda} f_X(s_\alpha(PA_X))$ would not give this result, since $f_Y(1, 0, 0, 0, \dots)$, $f_Y(1, 1, 0, 0, \dots)$, $f_Y(1, 1, 1, 0, \dots)$, \dots are all 0. The limit of a sequence of 0s is 0. So with this alternative we would undesirably predict $s_\omega(Y) = 0$.

that whenever $X \prec Y$, X is independent of the value of Y : $F_X(\dots, y, \dots) = F_X(\dots, y', \dots)$ for all values y, y' of Y . Recursive structural causal models can be represented by directed acyclic graphs. Also, in recursive structural causal models each assignment of values to the exogenous variables uniquely determines the value of every variable.

If a structural causal model is recursive, then when we follow the possible paths of its dynamic interpretation without changing the values of the exogenous variables, every variable eventually reaches its value according to the model's structural equations, and stays at that value forevermore. Formally, for any path (s_0, s_1, \dots) and variable X , let us say that $X = x$ is *eventually true* at the path just in case there is some s_t such that $s_{t'}(X) = x$ for all $t' \geq t$. Then we have the following fact.

6.5.3. PROPOSITION. *Let M be a recursive structural causal model and P_{fixExo}^+ the set of possible paths of its dynamic interpretation with the exogenous variables fixed. Let $\vec{U} = \vec{u}$ be a setting of the exogenous variables and (\vec{u}, \vec{v}) the unique setting of the variables containing \vec{u} that satisfies the structural equations of M . Let p be a path in P_{fixExo}^+ where $\vec{U} = \vec{u}$ holds at every state of p .*

For any variable X , $X = x$ is eventually true at p if and only if $X = x$ according to (\vec{u}, \vec{v}) .

Hence for any Boolean combination of assignments of values to variables φ , φ is eventually true at p if and only if φ is true according to (\vec{u}, \vec{v}) .

PROOF. Let M be a recursive structural causal model. We may assume that X is a parent of Y just in case Y depends on X , that is, $f_Y(\dots, x, \dots) \neq f_Y(\dots, x', \dots)$ for some values x, x' of X (if this does not hold, then we can equivalently make it hold by removing the variables that Y does not depend on from the domain of f_Y , for each variable Y). Pick any $(s_0, s_1, \dots) \in P_{fixExo}^+$ such that $s_0(\vec{U}) = \vec{u}$.

Let the *depth* of X , denoted $d(X)$, be the least ordinal that is greater than the depth of all of X 's parents: $d(X) = \min\{\alpha \in Ord : \alpha > d(Y) \text{ for all } Y \in PA_X\}$. (This is well-defined since M is recursive.) We show by transfinite induction on t that for any variable X with depth $d(X) \leq t$ we have $s_t(X) = x$, where x is the value X receives in (\vec{u}, \vec{v}) . From this the (\Leftarrow) direction will follow immediately.

In what follows, for any variables (e.g. X, Y, PA_X) we will let lower-case letters (e.g. x, y, pa_X) denote the value they receive in (\vec{u}, \vec{v}) .

Base case. If $t = 0$ then for any variable X with $d(X) \leq 0$, $d(X) = 0$, so X has no parents, i.e. it is exogenous. By assumption $s_0(\vec{U}) = \vec{u}$, so $s_t(X) = x$.

Successor case. Suppose $t = \alpha + 1$. Pick any variable X with depth $d(X) \leq t$. If X is exogenous then since we fix the exogenous variables, $s_t(X) = s_0(X) = x$. So suppose X is endogenous and pick any parent Y of X . By definition, Y has lower depth than X . Then $d(Y) \leq \alpha < t$, so by induction hypothesis, $s_\alpha(Y) = y$. Since Y was arbitrary, $s_\alpha(PA_X) = pa_X$, where pa_X is the value X 's parents receive in (\vec{u}, \vec{v}) . By definition of the dynamic interpretation $s_{\alpha+1}(X) = f_X(s_\alpha(PA_X)) = f_X(pa_X) = x$.

Limit case. Suppose $t = \lambda$ for some limit ordinal λ . Pick any variable X with depth $d(X) \leq t$. If X is exogenous then as before, $s_t(X) = x$. So suppose X is endogenous and pick any parent Y of X . By induction hypothesis, for all α with $d(Y) \leq \alpha < \lambda$, $s_\alpha(Y) = y$. Hence $\lim_{\alpha < \lambda} s_\alpha(Y) = y$. Since Y was arbitrary, $\lim_{\alpha < \lambda} s_\alpha(PA_X) = pa_X$. By definition of the dynamic interpretation, $s_\lambda(X) = f_X(\lim_{\alpha < \lambda} s_\alpha(PA_X)) = f_X(pa_X) = x$.

(\Rightarrow) Suppose $X = x$ is eventually true at (s_0, s_1, \dots) . Then for some time t , $s_t(X) = x$ for all $t' \geq t$. Suppose for reductio that $X \neq x$ according to (\vec{u}, \vec{v}) . Then $X = x'$ according to (\vec{u}, \vec{v}) for some $x' \neq x$. By the (\Leftarrow) direction, $X = x'$ is eventually true at p , i.e. there is a time t^* such that $s_{t'}(X) = x'$ for all $t' \geq t^*$. Since paths are linear, $t \geq t^*$ or $t^* \geq t$. If $t \geq t^*$ then $s_t(X) = x'$, contradicting $s_t(X) = x$. And if $t^* \geq t$ then $s_{t^*}(X) = x$, contradicting $s_{t^*}(X) = x'$.

For any Boolean combination of assignments of values to variables φ – i.e. any sentence in the language generated by $\varphi ::= X = x \mid \neg\varphi \mid \varphi \wedge \varphi \mid \varphi \vee \varphi$ – let the depth of φ , $d(\varphi)$, be the maximum depth of the variables appearing in φ .

We now show by induction on the complexity of φ that for any φ and ordinal $t \geq d(\varphi)$, $s_t(\varphi) = x$, where $x \in \{0, 1\}$ is the truth value of φ in (\vec{u}, \vec{v}) . We have just shown the atomic case.

Negation. Suppose φ is $\neg\psi$. Pick any ordinal $t \geq d(\varphi)$. Since φ and ψ contain the same variables, $d(\varphi) = d(\psi)$, so $t \geq d(\psi)$. Then by induction hypothesis, $s_t(\psi) = y$, where y is the truth value of ψ in (\vec{u}, \vec{v}) . Then $s_t(\neg\psi) = 1 - y$, which is also the truth value of $\neg\psi$ in (\vec{u}, \vec{v}) .

Conjunction. Suppose φ is $\psi \wedge \chi$. Pick any ordinal $t \geq d(\varphi)$. Then $d(\varphi) = \max\{d(\psi), d(\chi)\}$, so $t \geq d(\psi)$ and $t \geq d(\chi)$. Then by induction hypothesis, $s_t(\psi) = y$ and $s_t(\chi) = z$, where y and z are the truth values of ψ and χ , respectively, in (\vec{u}, \vec{v}) . Then $s_t(\psi \wedge \chi) = \min\{y, z\}$, which is the truth value of $\psi \wedge \chi$ in (\vec{u}, \vec{v}) . The disjunction case is similar.

(\Rightarrow) Suppose φ is not true at (\vec{u}, \vec{v}) . Then $\neg\varphi$ is true at (\vec{u}, \vec{v}) , so by the (\Leftarrow) direction, $\neg\varphi$ is eventually true at p . Then similar to the (\Rightarrow) case above, φ is not eventually true at p . \square

6.6 Interventions as sufficiency claims

One of the central uses of structural causal models is determine the effects of interventions. The goal of this section is to show that interventions can be seen as a special case of sufficiency claims, given the semantics of sufficiency we proposed in chapter 3.

The relationship between interventions and our framework is especially relevant since we have used our framework to provide a semantics of *would*-conditionals, while there are also a number of approaches to the semantics of *would*-conditionals

using structural causal models.²⁶ For example, Pearl (2009: chapter 7) presents a semantics of counterfactuals in terms of interventions, whereby a *would*-conditional is true just in case its consequent is true after an intervention to make its antecedent true.

An intervention is an operation on structural causal models where some variables are forced to take on certain values independently of the structural equations. More precisely, where $M = (U, V, R, F)$ is a structural causal model and \vec{X} a set of variables, an *intervention* on M to set $\vec{X} = \vec{x}$ results in the model $M_{\vec{X}=\vec{x}} = (U, V, R, F')$, where for every variable X in \vec{X} we replace the equation for X in F' with $X = x$, where x is the value X takes in \vec{x} , and the equation for every variable not in \vec{X} is unchanged ($f'_Y = f_Y$ for all $f'_Y \in F'$ with Y not in \vec{X}). For example, Figure 6.33 illustrates the effect of intervening to set $Y = y$.



Figure 6.33: A chain (left) and the result of intervening to set $Y = y$ (right).

Let us now state Pearl's semantics for *would*-conditionals. Following Halpern and Pearl (2005), for simplicity we will restrict attention to recursive structural causal models. Where M is a recursive structural causal model, let a *context* for M be an assignment \vec{u} of values to exogenous variables. Moreover, let $\vec{X} = \vec{x}$ be an assignment of values to some variables and φ be a Boolean combination of variable assignments. Pearl proposes that the *would*-conditional “if X had value x then φ would hold” is true at model M and context \vec{u} just in case φ is true at $M_{\vec{X}=\vec{x}}, \vec{u}$. Following Halpern and Pearl (2005:852) we symbolise this counterfactual claim as $[\vec{X} = \vec{x}]\varphi$. Then Pearl's interventionist semantics of counterfactuals is given by defining that $M, \vec{u} \models [\vec{X} = \vec{x}]\varphi$ if and only if $M_{\vec{X}=\vec{x}}, \vec{u} \models \varphi$.

Let us now briefly review our analysis of sufficiency from chapter 3. We defined that sentence A is sufficient for sentence C just in case C is true at every world in the modal horizon where A is true. To determine the modal horizon we pick a moment t at which to imagine a sudden change and consider all the moments t' that agree on everything A is not about (i.e. the A -variants of t). The modal horizon is the set of worlds that result from taking the nomically possible futures of an A -variant of t and sticking on the actual past (see Figure 3.22).

(3) For any world w , sentence A , set of worlds P and moment t , define

$$mh_{P,t}(w, A) := \{w_{\prec t} \frown w'_{\succeq t'} : t' \text{ is an } A\text{-variant of } t, t' \in w' \text{ and } w' \in P\}$$

²⁶Examples include Pearl (1995), Galles and Pearl (1998), Schulz (2007, 2011), Briggs (2012), Kaufmann (2013), Santorio (2014, 2019), and Ciardelli, Zhang, and Champollion (2018).

Finally, recall our semantics of sufficiency (\gg) and *would*-conditionals ($>$), repeated in (4).

- (4) Where $M = (S, \leq, \mathcal{A}, P, |\cdot|)$ is a model, w a world of M , t the intervention time and s the selection function,

$$\begin{aligned} M, w, t \models A \gg C & \quad \text{iff} & \quad mh_{P,t}(w, A) \cap |A| \subseteq |C| \\ M, w, t, s \models A > C & \quad \text{iff} & \quad s(w, mh_{P,t}(w, A) \cap |A|) \in |C| \end{aligned}$$

Our main contribution in this section is to show that we can see interventions in structural causal models as a special case of sufficiency. More precisely, given a structural causal model M , context \vec{u} for M and intervention $X = x$, we will construct a model M' in our framework, world w and intervention time t such that $[X = x]Y = y$ is true at M, \vec{u} just in case $X = x \gg Y = y$ is true in M' at w at t .

While we have spoken of interventions as a special case of sufficiency, the point applies to *would*-conditionals as well. Recall that the difference between sufficiency and *would*-conditionals is one of modal force: sufficiency quantifies universally over a set worlds while *would*-conditionals select a single world from this set. This difference will turn out not to matter when we restrict attention to recursive structural causal models (as we do in this section), since in recursive structural causal models the outcomes are interventions are unique. That is, given a recursive structural causal model M and a context \vec{u} for M , each intervention $\vec{X} = \vec{x}$ returns a single assignment of values to the variables; namely, $(\vec{u}, \vec{x}, \vec{v})$ where \vec{v} is the unique setting of the exogenous variables satisfying the equations in $M_{\vec{X}=\vec{x}}$.

Let $M = (U, V, R, F)$ be a recursive structural causal model and \vec{X} some variables of M . We construct a model in our framework, what we may call the *complete dynamic interpretation* of M , $M' = (S_M, \leq_M, \mathcal{A}_M, P_{M, \vec{X}=\vec{x}}, |\cdot|_M)$ as follows.

- **The state space.** A *state* of M is an assignment of values to some of M 's variables. A state s is *part* of state s' just in case s assigns the same value as s' to all of the variables that receive a value in s . For example, the state that assigns $Y = y$ is part of the state that assigns $(Y = y, Z = z)$.

$$\begin{aligned} S_M &= \{s : \vec{Y} \rightarrow R(\vec{Y}) \mid \vec{Y} \text{ is a nonempty subset of } U \cup V\} \\ \leq_M &= \subseteq \end{aligned}$$

- **Aboutness.** A sentence $\vec{Y} = \vec{y}$ is *about* state s just in case s assigns a value to all and only the variables in \vec{Y} .

$$\mathcal{A}_M = \{(\vec{Y} = \vec{y}, s) \mid s : \vec{Y} \rightarrow R(\vec{Y})\}$$

- **Nomic possibility.** Recall that *world* of M is a sequence of states of M , (s_0, s_1, \dots) of M . A world of M is nomically possible just in case it is possible in the dynamic interpretation of $M_{\vec{X}=\vec{x}}$ where we fix the exogenous variables. So where (P^+, P^-) is the dynamic interpretation of $M_{\vec{X}=\vec{x}}$,

$$P_{M, \vec{X}=\vec{x}} = P_{fixExo}^+$$

- **The interpretation function.** For any variable Y and value y , $Y = y$ is true at a world just in case Y eventually settles on value y at that world. That is, $Y = y$ is true at (s_0, s_1, \dots) just in case there is a time t such that $s_{t'}(Y) = y$ for all $t' \geq t$.²⁷

$$|Y = y|_M = \{(s_0, s_1, \dots) \in \text{worlds}(M) : \text{for some } t, s_{t'}(Y) = y \text{ for all } t' \geq t\}$$

The idea behind this definition is to give the structural equations enough time to work out the value of each variable. We extend the interpretation function to Boolean combinations of variable assignments in the usual way; for example, $Y = y \vee Z = z$ is true at a world just in case $Y = y$ or $Z = z$ is true at the world, and so on.

All of the choices above are fairly natural. Under these constraints on the model, interventions fall out as a special case of sufficiency, as shown by the following theorem. We work out an example below.

6.6.1. THEOREM. *Let M be a recursive structural causal model, \vec{u} a setting of the exogenous variables, $\vec{X} = \vec{x}$ an assignment of values to some variables and φ a Boolean combination of assignments of values to variables. Let $M' = (S_M, \leq_M, \mathcal{A}_M, |\cdot|_M, P_{M, \vec{X}=\vec{x}})$ be defined as above. Let w be any world of M' and t any moment of w such that $t(\vec{U}) = \vec{u}$. Then*

$$M, \vec{u} \models [\vec{X} = \vec{x}] \varphi \quad \text{if and only if} \quad M', w, t \models \vec{X} = \vec{x} \gg \varphi.$$

PROOF. (\Rightarrow) Suppose $M_{\vec{X}=\vec{x}}, \vec{u} \models [\vec{X} = \vec{x}] \varphi$. For any assignment of values \vec{z} , let $\vec{z}_{-\vec{X}}$ be the restriction of \vec{z} to the values not in \vec{X} . Since M is recursive, there is a unique setting of the variables $(\vec{u}_{-\vec{X}}, \vec{x}, \vec{v}_{-\vec{X}})$ extending $\vec{u}_{-\vec{X}}$ that satisfies the structural equations of $M_{\vec{X}=\vec{x}}$. And as $M_{\vec{X}=\vec{x}}, \vec{u} \models [\vec{X} = \vec{x}] \varphi$, φ is true in $(\vec{u}_{-\vec{X}}, \vec{x}, \vec{v}_{-\vec{X}})$.

²⁷Note we do not define that $Y = y$ is true at a world w just in case it is true at the equilibrium states of w , where for any world $w = (s_t, s_1, \dots)$, we say s_t is an *equilibrium state* of w just in case $s_t(X) = s_{t'}(X)$ for all $X \in V$ and $t' \geq t$. Consider the structural causal model $X_0 \rightarrow X_1 \rightarrow X_2 \rightarrow \dots$ where for all $n \in \mathbb{N}$, $X_{n+1} = X_n$. Intuitively we do not need transfinite time to give this model a dynamic interpretation. Taking the worlds of this model to be of the form $(s_t)_{t \in \mathbb{N}}$, the world given by $(1, 0, 0, \dots), (1, 1, 0, \dots), (1, 1, 1, \dots), \dots$ does not have any equilibrium states, but every variable eventually settles on 1 in this world.

To show that $M', w, t \models \vec{X} = \vec{x} \gg \varphi$, by our semantics of sufficiency in (4) we have to show that $mh_{P_{M, \vec{X}=\vec{x}}, t}(w, \vec{X} = \vec{x}) \cap |\vec{X} = \vec{x}|_M \subseteq |\varphi|_M$. Pick any world $w' \in mh_{P_{M, \vec{X}=\vec{x}}, t}(w, \vec{X} = \vec{x}) \cap |\vec{X} = \vec{x}|_M$. By our definition of the modal horizon in (3), w' is of the form $w_{\leq t} \frown w'_{\geq t'}$ for some $(\vec{X} = \vec{x})$ -variant t' of t , $t' \in w'$ and $w' \in P_{M, \vec{X}=\vec{x}}$. We show that $w' \in |\varphi|_M$.

Recall from chapter 3 that for any state t' , t' is an $(\vec{X} = \vec{x})$ -variant of t just in case every part of t that does not overlap any state $\vec{X} = \vec{x}$ is about is part of t' . Let $s : \vec{U}_{-\vec{X}} \rightarrow R(\vec{U}_{-\vec{X}})$ be the state assigning $\vec{u}_{-\vec{X}}$ to the exogenous variables not in \vec{X} . By definition of the aboutness relation \mathcal{A}_M , the sentence $\vec{X} = \vec{x}$ is not about s , so s is part of t' . That is, $t'(\vec{U}_{-\vec{X}}) = \vec{u}_{-\vec{X}}$. By construction, w' is of the form (s_0, s_1, \dots) where $s_0 = t'$. Since $w' \in P_{M, \vec{X}=\vec{x}}$, w' is nomically possible according to the dynamic translation of $M_{\vec{X}=\vec{x}}$ where we fix the exogenous variables. Then as $s_0(\vec{U}_{-\vec{X}}) = \vec{u}_{-\vec{X}}$, also $s_1(\vec{U}_{-\vec{X}}) = \vec{u}_{-\vec{X}}$. And since the equation in $M_{\vec{X}=\vec{x}}$ for each X in \vec{X} is $X = x$, $s_1(X) = f_X(s_0(PA_X)) = x$. Thus $s_1(\vec{U}_{-\vec{X}}, \vec{X}) = (\vec{u}_{-\vec{X}}, \vec{x})$. That is, $(\vec{U}_{-\vec{X}}, \vec{X}) = (\vec{u}_{-\vec{X}}, \vec{x})$ is true at the first state of the path (s_1, s_2, \dots) . Also, this path is in P_{fixExo}^+ , and φ is true at $(\vec{u}_{-\vec{X}}, \vec{x}, \vec{v}_{-\vec{X}})$. Then by Proposition 6.5.3, φ is eventually true at (s_1, s_2, \dots) .

Note that for any path p , if φ is eventually true at p then φ is also eventually true at $p' \frown p$ for any path p' , where \frown denotes concatenation. Since $w' = w_{\leq t} \frown w'_{\geq t'} = w_{\leq t} \frown s_0 \frown (s_1, s_2, \dots)$, φ is eventually true at w' , that is, $w' \in |\varphi|_M$.

(\Leftarrow) Suppose $M_{\vec{X}=\vec{x}}, \vec{u} \not\models [\vec{X} = \vec{x}]\varphi$. Since M is recursive, the effects of interventions in M are unique, so $M_{\vec{X}=\vec{x}}, \vec{u} \models [\vec{X} = \vec{x}]\neg\varphi$. By the (\Rightarrow) direction, $M', w, t \models \vec{X} = \vec{x} \gg \neg\varphi$. By construction of M' , clearly $mh_{P_{M, \vec{X}=\vec{x}}, t}(w, \vec{X} = \vec{x}) \cap |\vec{X} = \vec{x}|_M$ is nonempty. So $M', w, t \not\models \vec{X} = \vec{x} \gg \varphi$. \square

Here is an example to illustrate the theorem. Consider two switches connected to a light. We may represent this system as a structural causal model, depicted in Figure 6.34 together with its dynamic interpretation with the exogenous variables fixed. (As usual, a world is nomically possible just in case it is a directed path through the figure). Our construction does not take the dynamic interpretation of this model; rather, it takes the dynamic interpretation of the post-intervention model, still with the exogenous variables fixed, shown in Figure 6.35.

Say we are in a context where switch 1 is down, switch 2 is up, and the light is off. Intervening to set switch 2 down results in a state where the light is on: $M, \vec{u} \models [S_2 = 1]L = 1$. We replicate this with sufficiency using the construction of the model M' in our framework. Pick any moment t where the exogenous variables have values \vec{u} ; say t assigns $(S_1 = 1, S_2 = 0, L = 0)$. By our construction of aboutness in M' , the sentence $S_2 = 1$ is not about the states $S_1 = 1$ or $L = 0$. So the $(S_2 = 1)$ -variants of t are $(S_1 = 1, S_2 = 0, L = 0)$ and $(S_1 = 1, S_2 = 1, L = 0)$. We then consider all the worlds containing these

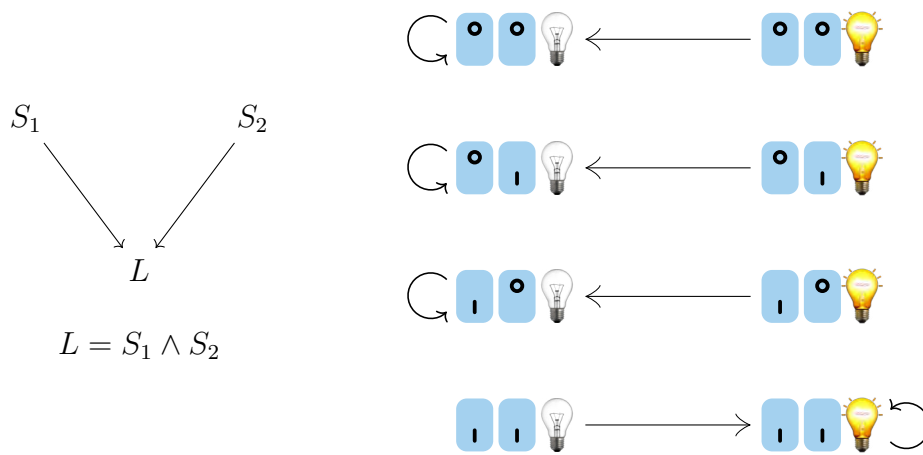


Figure 6.34: An AND gate as a structural causal model (left) and its dynamic interpretation with the exogenous variables fixed (right).

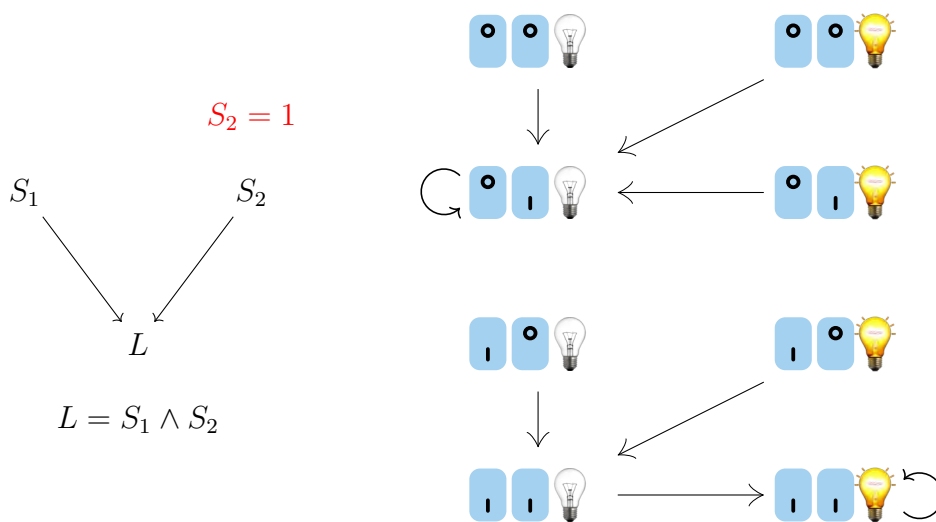


Figure 6.35: The intervened model $M_{S_2=1}$ (left) and its dynamic interpretation with the exogenous variables fixed (right).

states that are nomically possible in the dynamic interpretation of $M_{S_2=1}$ with the exogenous variables fixed. As Figure 6.35 shows, there are two such worlds. At every world in the modal horizon, the light eventually turns on and stays on

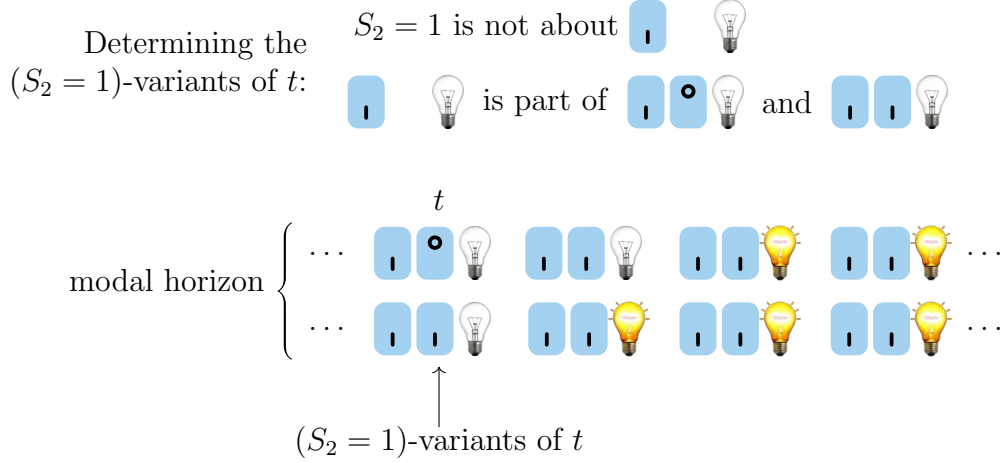


Figure 6.36: Determining that $S_2 = 1$ is sufficient for $L = 1$ at M' .

forever. So $mh_{P,t}(w, S_2 = 1) \subseteq |L = 1|_M$. A fortiori, $mh_{P,t}(w, S_2 = 1) \cap |S_2 = 1|_M \subseteq |L = 1|_M$, so $M', w, t \models S_2 = 1 \gg L = 1$. This shows that the intervention claim $[S_2 = 1]L = 1$ in the structural causal model M falls out as a special case of the sufficiency claim $S_1 = 1 \gg L = 1$; namely, as sufficiency in the model M' we constructed in our framework.

6.6.1 From sufficiency to *would*-conditionals

Since interventions in recursive structural causal models are unique, we can use Theorem 6.6.1 to also show that interventions are also a special case of our semantics of *would*-conditionals: for any selection function s we have

$$M_{\vec{X}=\vec{x}}, \vec{u} \models [\vec{X} = \vec{x}] \varphi \quad \text{if and only if} \quad M', w, t, s \models \vec{X} = \vec{x} > \varphi$$

where M , \vec{u} , $\vec{X} = \vec{x}$, φ and M' are given as stated in Theorem 6.6.1. The left-to-right direction follows since sufficiency entails the corresponding *would*-conditional.

$$\begin{aligned} & M_{\vec{X}=\vec{x}}, \vec{u} \models [\vec{X} = \vec{x}] \varphi \\ \Rightarrow & M', w, t \models \vec{X} = \vec{x} \gg \varphi && \text{(Theorem 6.6.1)} \\ \Rightarrow & M', w, t, s \models \vec{X} = \vec{x} > \varphi && (> \text{ entails } \gg) \end{aligned}$$

The right-to-left direction follows, as above, from the fact that in recursive structural causal models the effects of interventions are unique.

$$\begin{aligned}
& M_{\vec{X}=\vec{x}}, \vec{u} \not\models [\vec{X} = \vec{x}]\varphi \\
\Rightarrow & M_{\vec{X}=\vec{x}}, \vec{u} \models [\vec{X} = \vec{x}]\neg\varphi && (M \text{ is recursive}) \\
\Rightarrow & M', w, t \models \vec{X} = \vec{x} \gg \neg\varphi && (\text{Theorem 6.6.1}) \\
\Rightarrow & M', w, t, s \models \vec{X} = \vec{x} > \neg\varphi && (\gg \text{ entails } >) \\
\Rightarrow & M', w, t, s \not\models \vec{X} = \vec{x} > \varphi && (\text{The modal horizon is nonempty})
\end{aligned}$$

6.6.2 Comparing interventions and *would*-conditionals

We have seen that interventions claims in structural causal models fall out as a special case of sufficiency claims and *would*-conditionals in our framework; namely, when we construct our model M' according to the constraints above. In general, our framework does not have to adhere to these. Here are three cases showing how our framework goes beyond these constraints.

Simplification 1. The state space of M' has a variable structure. In the model M' we constructed from a structural causal model, the state space is generated by a set of variables. In section 3.4.3 we saw some examples of state spaces without a variable structure (e.g. the state space representing an object's colour and material) and showed that our framework makes reasonable predictions in these cases.

Simplification 2. The aboutness relation of M' ignores nomic relevance. Consider the following scenario, due to Hiddleston (2005). If the cannon is lit, there is a simultaneous flash and bang. Actually, the cannon was not lit, there was no flash, and no bang.

(5) If there had been a flash, there would have been a bang.

Hiddleston observes that (5) is intuitively acceptable. Figure 6.37 presents a plausible structural causal model and context representing Hiddleston's scenario, as well as the result of intervening to set there to be a flash.

As we see, this model incorrectly predicts (5) to be false.²⁸

In contrast, under plausible assumptions our semantics of *would*-conditionals from chapter 3 predicts (5) to be true. In section 3.7.1 we discussed the relationship between nomic relevance and aboutness. A plausible principle is the *minimal nomic relevance constraint*: for any sentence A and situation s , if s is minimally

²⁸Though for a recent account of backtracking conditionals within structural causal models, see Kügelgen, Mohamed, and Beckers (2022).



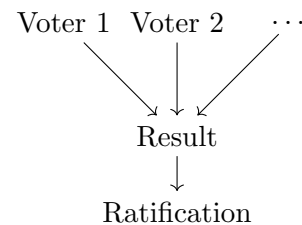
Figure 6.37

nomically relevant to A then A is about s .²⁹ The state of the cannon not being lit is minimally nomically relevant to *There was a flash*, since in every nomically possible world containing this state the sentence is false. So by the minimal nomic relevance constraint, *There was a flash* is about the state of the cannon being lit. Then when we imagine there being a flash, we remove the state of the cannot not being lit. In every nomically possible world where there is a flash, the cannon was lit and there is a bang. Thus the minimal nomic relevant constraint, paired with our analysis of *would*-conditionals, correctly predicts (5) to be true.

Simplification 3. The interpretation function of M' ignores utterance time. Sufficiency and interventions have a different relationship with time. In a sufficiency claim $X = x \gg \varphi$, we evaluate φ at worlds – sequences of states extended through time. This follows standard practice in possible-worlds semantics, which means we can take existing work on the semantics of tense and plug it directly into the present framework, via the model’s interpretation function.

In contrast, with an intervention $[X = x]\varphi$, we evaluate φ at an assignment of values to variables. Any notion of time must be introduced by the variables themselves (say, by indexing variables with times, as in dynamic Bayesian networks; see Dagum, Galper, and Horvitz 1992). For a semantics of *would*-conditionals using structural causal models to interact with existing work in natural language semantics we have to reconstruct the models of natural language semantics using the variables.

One feature of the interpretation function in natural language semantics which is absent from structural causal models is the use of an evaluation time. To illustrate, imagine a country where the parliament votes on bills: if a bill passes, it is signed into law on January 1st of the next year. We may represent this with the model on the right, where *Voter i* represents how i



²⁹Recall that a situation s is *nomically relevant* to a sentence A just in case A is true at every nomically possible world containing s , or A is false at every nomically possible world containing s , or A is true at every nomically possible world not containing s , or A is false at every nomically possible world not containing s . And s is *minimally nomically relevant* to A just in case s is nomically relevant to A and no proper part of s is nomically relevant to A .

votes, *Result* whether the bill receives a majority and *Ratification* whether it is signed into law.

Suppose it is September. The parliament has just voted on a bill which failed by one vote. Alice did not vote for the bill. Consider (6), uttered in September.

- (6) a. If Alice had voted for the bill, it would be law.
- b. If Alice had voted for the bill, it would become law.

(6a) is intuitively false, implying that the bill would be law at the time of utterance, while (6b) is intuitively true, implying that the bill would be law in the future. This is an instance of a general pattern in tense-modal interaction.³⁰ Modals, such as *might* and *would*, interact differently with statives (e.g. *be law*) and eventives (e.g. *become law*), as illustrated in (7) from Condoravdi (2002).

- (7) a. He might be sick.
- b. He might get sick.

Condoravdi (2002) accounts for the contrast by proposing that modals set their time of evaluation to the interval beginning at utterance time and extending indefinitely into the future, and that eventives (e.g. *get sick*) require the described event to take place within the evaluation time. On the other hand, (6a) shows that statives under modals do not shift the evaluation time. This account of course requires appeal to an evaluation time, which the interpretation function $|\cdot|_M$ ignores, but in general the interpretation function $|\cdot|$ may take into account.

6.6.3 Probabilistic dynamic interpretations

We can also give a probabilistic version of the dynamic interpretation. This will allow us to give a dynamic interpretation of Bayesian networks, interpreting any Bayesian network as a discrete time Markov chain.

Let us first introduce some terminology. Where $G = (V, E)$ a directed acyclic graph, for any $X, Y \in V$ let us say that X is a *parent* of Y just in case $(X, Y) \in E$ and that Y is a *descendent* of X just in case there is a directed path from X to Y (that is, the descendent relation is the transitive closure of the parent relation).

6.6.2. DEFINITION (Bayesian network (Pearl 1988)). Let V be a set of variables, $G = (V, E)$ a directed acyclic graph and P a joint probability distribution over V . We say (V, E, P) satisfies the *Markov condition* just in case every variable is independent of its non-descendants conditional on its parents:

$$P(x \mid pa_X, nd_X) = P(x \mid pa_X)$$

for every value x of X , value pa_X of X 's parents and value nd_X of X 's non-descendants. (V, E, P) satisfies the *minimality condition* just in case for no $E' \subsetneq E$

³⁰For an overview of tense-modal interaction see Fălăuş and Laca (2020).

E does (V, E', P) satisfy the Markov condition. Lastly, (V, E, P) is a *Bayesian network* just in case it satisfies the Markov and minimality conditions.

If P is Markov relative to G , then by the chain rule we can express the probability of a variable assignment in terms of the probabilities of its parents like so:

$$P(x_0, x_1, \dots) = \prod_i P(x_i \mid pa_{X_i})$$

Here is a slightly different way to write this equation. Where $s : V \rightarrow R(V)$ is an assignment of values to the variables, we may write it as:

$$P(s) = \prod_i P(s(X_i) \mid s(PA_{X_i}))$$

Notice that for each variable X the equation uses the value of its parents at a state to determine the value of the variable *at that same state*. Sometimes, however, we wish to reason across states. For example, we often wish to determine the probability of a future event in light of current information. If we take states to represent the value of the variables at a given moment in time, the above expression gives us a way to calculate the probability that we are currently in a particular state, but does not give us a way to calculate future probabilities from present information.

Suppose we interpret the graph of a Bayesian network as expressing dependence in time, whereby the probability of a variable having a certain value at one moment is determined by the values of its parents at the previous moment. Then we can determine the probability of one state s transitioning into a state t as follows.

$$P_s(t) = \prod_i P(t(X_i) \mid s(pa_{X_i}))$$

Figure 6.38 illustrates this result of this equation for a Bayesian network $X \rightarrow Y$ with two binary variables.

One can extend P_s to sets of states by putting $P_s(T) = \sum_{t \in T} P_s(t)$ for any set of states T . It turns out that P_s , defined in this way, is still a probability distribution, which we show now.

6.6.3. FACT. Let $V = \{X_1, \dots, X_n\}$ be a set of variables, each with range $R(X_i)$ and let $G = (V, E)$ a directed acyclic graph. Let $S = \{s \mid s : V \rightarrow R(V)\}$ be the set of assignments of values to variables, and $P : S \rightarrow [0, 1]$ be probability distribution that is Markov relative to G . For any $s, t \in S$, define $P_s(t) = \prod_i P(t(X_i) \mid s(PA_{X_i}))$ and $P_s(T) = \sum_{t \in T} P_s(t)$ for any set of states T . Then for any $s \in S$, $P_s : \mathcal{P}(S) \rightarrow [0, 1]$ is a probability distribution.

PROOF. Clearly $P_s(T)$ is non-negative for any set of states T .

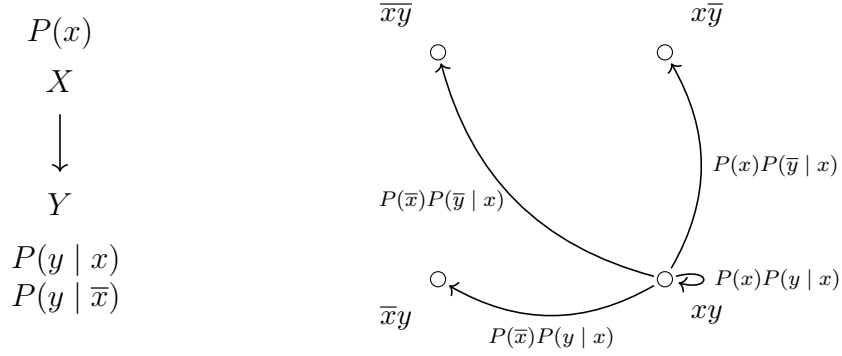


Figure 6.38: Illustrating the dynamic interpretation of Bayesian networks

We show $P_s(S) = 1$ by induction on the number of variables in the graph. *Base case.* Let $n = 1$. Then $P_s(S) = \sum_{t \in S} P_s(t) = \sum_{t \in S} P(t(X) \mid \emptyset) = \sum_x P(x) = 1$.

Induction step. Since n is finite and the graph is acyclic, there is a variable X_n without any children. Let $V' = V \setminus \{X_n\}$ and $E' = E \cap (V' \times V')$ and P' the restriction of P to V' ; that is, $S' = \{s \mid s : V' \rightarrow R(V')\}$ and $P' : S' \rightarrow [0, 1]$ where $P'(x_0, \dots, x_{n-1}) = P(x_0, \dots, x_{n-1})$ for all $x_0, \dots, x_{n-1} \in R(V')$. And since X_n does not have any children, $s(pa'_{X_i}) = s(pa_{X_i})$ for any $X_i \in V'$, where pa'_{X_i} is the set of X_i 's children in G' . Since (V, E, P) is a Bayesian network, clearly (V', E', P') is too. Then by induction hypothesis, $\sum_{t \in S'} P_s(t) = \sum_{t \in S'} \prod_{i=1}^{n-1} P(t(X_i) \mid s(pa_{X_i})) = 1$. Thus

$$\begin{aligned}
 \sum_{t \in S} P_s(t) &= \sum_{t \in S} \prod_{i=1}^n P(t(X_i) \mid s(pa_{X_i})) \\
 &= \sum_{x_n} \left(P(x_n \mid s(pa_{X_n})) \cdot \sum_{t \in S'} \prod_{i=1}^{n-1} P(t(X_i) \mid s(pa_{X_i})) \right) \\
 &= \sum_{x_n} P(x_n \mid s(pa_{X_n})) \cdot 1 \\
 &= 1
 \end{aligned}$$

Additivity follows by definition: for any disjoint sets of states T, U , $P_s(T \cup U) = \sum_{t \in T \cup U} P_s(t) = (\sum_{t \in T} P_s(t)) + (\sum_{u \in U} P_s(u)) = P_s(T) + P_s(U)$. \square

6.7 Dense causal chains

The dynamic interpretation provides a way to translate structural causal models and Bayesian networks into the present modelling framework. The question naturally arises whether one can go the other way. Is every model in our framework

the dynamic interpretation of a structural causal model or Bayesian network? In section 6.6.2 we have already seen that this is not the case. Our modelling framework enjoys greater generality: it works with state spaces that do not have a variable structure, allows aboutness relations that predict backtracking readings of conditionals, and adopts a more general interpretation function from possible worlds semantics (rather than an interpretation function based on eventual truth). So far we have not discussed this issue for nomic possibility. Does our framework have a more general notion of nomic possibility than that provided by structural causal models and Bayesian networks?

This question is relevant since some have proposed that structural causal models and Bayesian networks provide a fully general framework in which to represent causal dependence. In the provocatively titled paper ‘Bayesian Nets Are All There Is To Causal Dependence’, Wolfgang Spohn argues that “In the final analysis it is the all-embracing Bayesian net representing the whole of reality which decides about how the causal dependencies actually are” (Spohn 2001, reprinted as Spohn 2009). Spohn later adds the caveat that

it is not wholly true that Bayesian nets exhaust all there is to the notion of causal dependence. I have hardly addressed the relation between time and causation and not at all the relation between space and causation, and both add considerably to the notion of causal dependence, i.e., to how the all-embracing Bayesian net has to look in the final analysis.

Though here Spohn still seems to think that the final analysis will be expressed in terms of an all-embracing Bayesian net.

In this section we show that this is mistaken. Our framework has a strictly more general notion of nomic possibility than that provided by structural causal models and Bayesian networks. There are some scenarios whose nomic possibilities can be expressed in our framework but cannot be expressed by any structural causal model or Bayesian network.

The expressive limitation we will discuss has been present right from the birth of causal inference in statistics. The first sentence of Sewall Wright’s landmark paper ‘Correlation and Causation’ in 1921 reads:

The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated.

(Wright 1921:557)

Notice Wright’s focus on *direct* influence. The focus lives on today. For example, Greenland and Pearl (2011:208–09) write,

In causal diagrams, an arrow represents a “direct effect” of the parent on the child, although this effect is direct only relative to a certain

level of abstraction, in that the graph omits any variables that might mediate the effect.

Notice that Wright, Greenland and Pearl focus on *direct* influence. This focus is perfectly legitimate for the discrete systems that concerned Wright, such as the population dynamics of guinea pigs. But given this focus on *direct* dependence, expressive limitations are inevitable. For instance, when one represents space and time as dense – in the sense that between any two points there is a third – it is natural to suppose that the effect of any point on another is mediated through infinitely many points between them. In this case the causal relationships are always mediated and are therefore, fundamentally, *indirect*. Causal relationships in dense systems cannot be analysed in terms of direct dependence, at least not without modifying the scenario abstracting the density away.

The rest of this chapter is dedicated to showing that this intuitive argument translates into impossibility results. That is, we define what it would mean for a structural causal model or Bayesian network to represent dense causal dependence, and prove that no structural causal model or Bayesian network does so. Though before we move on, let us briefly address the issue of whether there are actual cases of dense causal dependence.

6.7.1 The reality of dense dependence

Loosely, to say that a causal relation is dense is to say that the causal influence of one event on another is mediated by infinitely many events located between them.³¹

Our intuitive picture of the world is filled with dense causal dependence. This is a consequence of two everyday assumptions. First, that the dimensions of space and time are *dense*, that is, between any two points in space or time there is a point between them. Second, that causal dependence is *local*: any causal influence from a cause to its effect passes continuously through the space and time between them. These ideas also appear in physics, where dimensions are represented as real numbers under their usual ordering – which is dense – and where there is broad support for the principle of locality, which says, intuitively, that any point in space is influenced only by its immediate surroundings.³² Together, density and locality imply the existence of dense causal dependence.

³¹The property of dense causal dependence will be formulated precisely later in the chapter.

³²The principle of locality is satisfied in classical mechanics, classical electrodynamics, and relativity, though is taken to be inconsistent with a realist view of quantum mechanics assuming Bell's inequality. While locality is taken by many to be inconsistent with a realist view of quantum mechanics given Bell's inequality (though this is disputed; see Goldstein et al. 2011), the vast majority of phenomena studied within quantum mechanics do operate locally, and are represented in dense dimensions. One would expect causal models to be able to represent the causal relations that hold in such systems. Nonetheless, one does not need locality to be universally valid to demonstrate the expressive limitation of Bayesian networks and structural causal models. The proofs to follow show that Bayesian networks and structural causal models

Of course physicists may well find reasons to believe that spacetime is not dense after all (the Planck length could turn out to be the shortest possible length), or that locality is violated (as some believe in light of Bell's results). But there is an argument that regardless of whether the world actually contains any dense causal dependence, causal models should be able to represent dense causal dependence. The argument is that causal models should be able to capture all instances of causal *reasoning*, even about nonactual possibilities. It is conceivable that the world contains instances of dense causal dependence; we seem, moreover, perfectly capable of reasoning about causation in such a world. If one believes that causal models should be able to represent our reasoning about dense causal dependence, the expressive limitation of Bayesian networks and structural causal models stands.

6.7.2 An example: a light system

To ground our discussion to come, let us fix a particular example of dense dependence. Consider a torch that produces a beam of light. Since light moves, whether a particular point in space is illuminated at a given time depends on the points around it.

For instance, if a point x in space is illuminated at time t then for any $y > x$, y will be illuminated at $t + \frac{y-x}{c}$, where c is the speed of light and $y - x$ the distance between x and y along the trajectory of the beam. The situation is illustrated in Figure 6.39, where the light turns on at time t and l is the point in space where the beam begins.

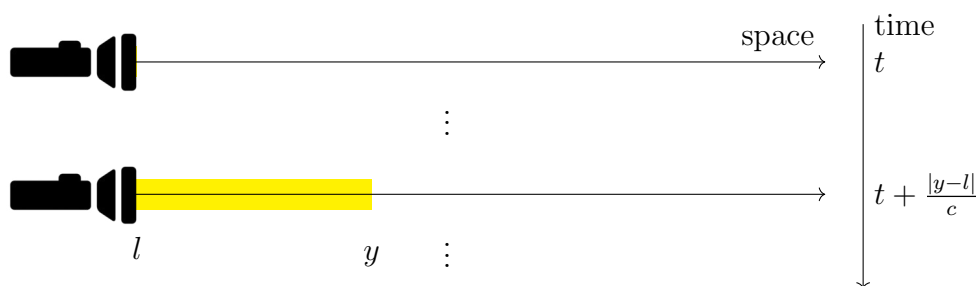


Figure 6.39: Turning on the light at time t

The dots between the lines are designed to show that space and time in the system are *dense*. The two states of the system depicted are just two snapshots in the continuous flow of time, with the light from x reaching infinitely many points before reaching y .

cannot represent *any* instance of dense causal dependence. Thus, as long as dimensions are assumed to be dense and locality is not violated everywhere, we have an scenario whose causal structure cannot be represented by Bayesian nets or structural causal models.

The states of the system can be represented by binary variables. For each point in space x and point in time t we need a variable representing whether the location x is light (1) or dark (0) at that time. For convenience, we will assume that space in the system is one-dimensional, existing only along the trajectory of the beam. Then let T be a set representing points in time and let X be a set representing points in space (e.g. T and S could each be a copy of the real numbers). Our set of variables will be $\{x_t : x \in S, t \in T\}$ (we assume that the state of the torch is represented by the point $x = 0$).

6.7.3 Dense causal chains in the present framework

The present framework can straightforwardly represent the causal relations holding in this system. Let x and y be points in space with $y > x$, and t and t' times with $t' = t + \frac{y-x}{c}$. Suppose that x was lit at t and y was lit at t' . Consider:

- (8) a. Point x being lit at t caused y to be lit at t' .
 b. Point y was lit at t' because x was lit at t .

(8) are intuitively true. To account for this in the framework we have proposed we first need a model: a state space, aboutness relation, nomic possibility and interpretation function.

Model. The construction is straightforward. We create our state space in the usual way. For each time t , a state is an assignment $s_t : S' \rightarrow \{0, 1\}$ where $S' \subseteq S$. That is, each state s_t determines for some points in space whether or not they are lit at t . As before, state s is part of state s' just in case $s \subseteq s'$. We assume that the sentence $x_t = 1$ is about the state assigning s_t that assigns $s_t(x) = 1$ and does not determine the value of any other points (the sentence may also be about other states, but this is all we need).

A world is a sequence of maximal states: assignments of 0 or 1 to each point in space. A world is nomically possible just in case for any points in space x and y and times t and t' such that the light at x at t can reach y at t' (i.e. $t' = t + (y-x)/c$), if x is lit at t , y is lit at t' , and if x is not lit at t , y is not lit at t' .

$$W = \{(s_t)_{t \in T} \mid s_t : S \rightarrow \{0, 1\} \text{ for all } t \in T\}$$

$$P = \{(s_t)_{t \in T} \in W \mid s_t(x) = s_{t'}(y) \text{ for all } x, y \in S \text{ and } t, t' \in T \text{ with } t' = t + (y-x)/c\}$$

This is illustrated in Figure 6.40, where space lies on the horizontal axis and time flows from top to bottom.

Finally, we assume that $y_{t'} = 1$ is true at a world $(s_t)_{t \in T}$ just in case $s_{t'}(y) = 1$.

Meanings. Now that we have a model, we can apply our semantics of *cause* and *because* to predict the truth values of (8). We have to check the positive

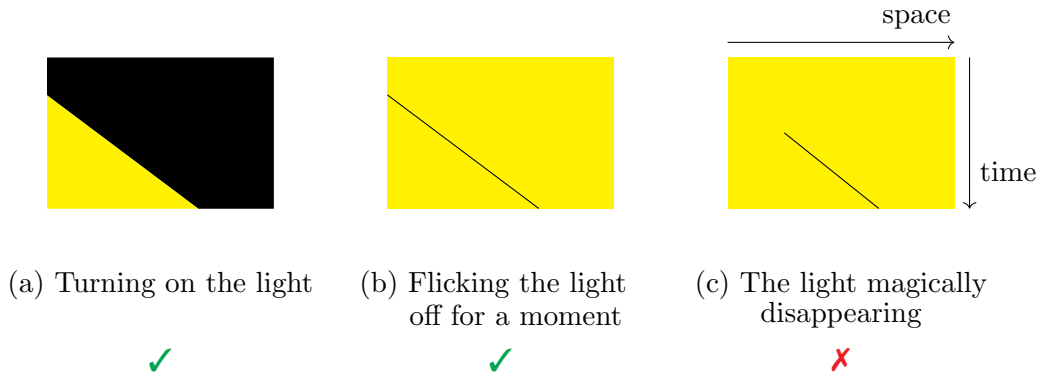


Figure 6.40: Two nomically possible worlds (✓) and one nomically impossible world (✗).

condition, that x being lit at t was sufficient for that to produce y to be lit at t' , and the negative condition, that x not being lit at t is not sufficient for that to produce y to be lit at t' .

The positive condition. There is only one ‘way’ for x to be lit at t , and the system is deterministic, so in this case sufficiency reduces to mere truth: $x_t = 1$ was sufficient for that to produce $y_{t'} = 1$ just in case $x_t = 1$ indeed produced $y_{t'} = 1$. So let us show this: x being lit at t produced y to be lit at t' .

Recall our analysis of production from chapter 5. $x_t = 1$ produced $y_{t'} = 1$ just in case there is a chain of dependence from $x_t = 1$ to $y_{t'} = 1$; more precisely, just in case there is a chain of propositions, convex in time, that begins with x_t and ends with $y_{t'}$, where for each proposition on the chain $p_{t''}$, there is a previous one on the chain $q_{t'''}$ such that $\neg q_{t'''} is sufficient for $\neg p_{t''}$ for all t''' with $t''' \leq t'' \leq t'$, with t''' the intervention time.$

Consider the spacetime points on the chain between x_t and $y_{t'}$: our chain is $C = \{z_{t''} = 1 : t \leq t'' \leq t' \text{ and } t'' = z_{-x}/c\}$, as illustrated in Figure 6.41.

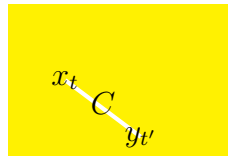


Figure 6.41

Pick any $(z_{t''} = 1)$ on the chain. For any $(u_{t'''} = 1)$ on the chain between $x_t = 1$ and $z_{t''} = 1$, $u_{t'''} = 0$ is sufficient for $z_{t''} = 0$, where t''' is the intervention time. The sentence $u_{t'''} = 0$ is about the state of point u being lit at t''' , so we remove that state and consider the state of the world at that time where u is not lit. In every nomically possible future of this moment, z is not lit at t'' (see Figure 6.42). Thus the dependence required by production is established.

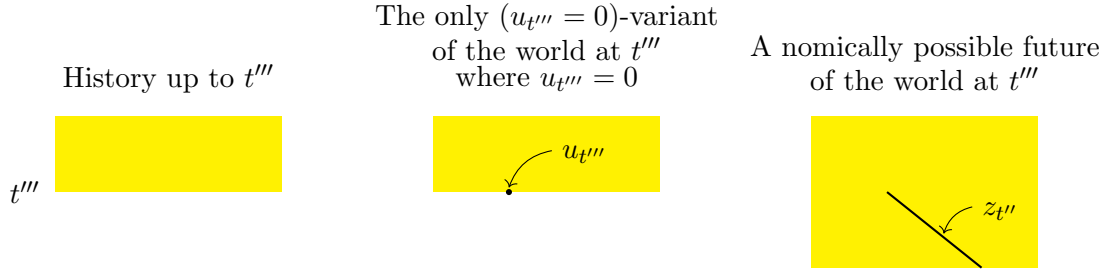


Figure 6.42: Illustrating that $u_{t'''} = 0$ is sufficient for $z_{t''} = 0$.

The negative condition. We have to show that $x_t = 0$ is not sufficient for $x_t = 0$ to produce $y_{t'} = 0$. It is enough to show that if x had not been lit at t , y would not have been lit at t' (this follows from factivity of production: if y is not lit at t' , then nothing produced it to be lit at t'). Given the nomic possibilities, this is clear: at every possible future of every $(x_t = 0)$ -variant of the world, $y_t = 0$. The present framework therefore correctly predicts (8) to be true.

Our model also captures the fact that the causal chain from any point x_t to $y_{t'}$ is dense: the influence from x_t to $y_{t'}$ must ‘pass through’ the points between them, in the sense that there is another point z and time t'' with $t < t'' < t'$ such that (9) holds.

- (9) If x were dark at t , then if z were still lit at t'' , y would still be lit at t'' .

We will formalise (9) as $x_t = 0 > (z_{t''} = 1 > y_t = 1)$. Interpreting sentences with nested conditionals such as (9) requires determining how the intervention time can evolve during the course of interpretation. There are two ways to go about this. The first is to put intervention times in the syntax (as free variables contributed, perhaps, by past tense morphology on the antecedent), which we can represent as $x_t = 0 >_i (z_{t''} = 1 >_{i'} y_t = 1)$, where i and i' are intervention times. The second option is to assume that the intervention time is a parameter of interpretation which, like parameters of interpretation in general, can evolve dynamically during the course of interpretation. For example, the utterance time may shift, as shown in (10a), and the modal horizon – that is, the set of possibilities at which we evaluate the consequent, which von Fintel (2001b) takes to be a parameter of interpretation that can evolve during the course of interpretation – may shift with nested conditionals, as shown in (10b).

- (10) a. It’s not your birthday yet, it’s your birthday NOW!
 b. If kangaroos had no tails, then if they used crutches, they would not topple over.

We need not decide between these two approaches here. The key takeaway is that interpreting nested conditionals can involve multiple intervention times. We will assume that when we interpret $x_t = 0$, we set the intervention time to t , and

when we interpret $z_{t''} = 1$, we set the intervention time to t'' . These seem to be natural assumptions. With them, we predict (9) to be true. At time t , we find the $(x_t = 0)$ -variant of the state of the system at time t where x is dark, and then take the possible futures of this moment. When we get to t'' , we take the $(z_{t''} = 1)$ -variant of the state of the system at time t'' where z is lit, and then take the possible futures of this moment. In each of these, y is lit at t' , so (9) comes out true. This reasoning is illustrated in Figure 6.43.

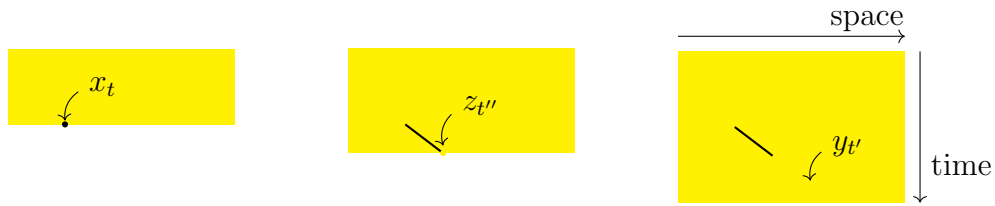


Figure 6.43: Illustrating how the present approach predicts (9) to be true.

Since x_t and $y_{t'}$ were arbitrary, what we have shown is that for any $x, y \in S$ and $t, t' \in T$ there is a $z \in S$ and $t'' \in T$ where (9) is true.

Let us now see how structural causal models fare with dense causal chains.

6.7.4 The impossibility of dense causal chains in structural causal models

It appears to be impossible to represent dense causal chains with structural causal models. Here is a simple argument for this conclusion. Representing dense causal chains in structural causal models would appear to require satisfying the following property, what we call dense dependence.

6.7.1. DEFINITION (Dense dependence). Let M be a structural causal model and Y a variable of M . We say dependence is *dense* at Y iff for every parent X of Y there is a parent Z of Y such that

$$f_Y(\dots, x, z, \dots) = f_Y(\dots, x', z, \dots)$$

for all values x, x' of X and value z of Z .

In other words, dependence is dense at Y just in case for every parent X of Y , there is another parent Z of Y such that holding Z fixed cuts off any dependence between X and Y .

Now recall that for any structural causal model M , variable Y and variables X , we say Y *depends on* X in M just in case there is a setting of Y 's parents such that changing the value of X results in a change in the value of Y :

$f_Y(\dots, x, \dots) \neq f_Y(\dots, x', \dots)$ for some values x, x' of X . We then have the following straightforward impossibility result.

6.7.2. PROPOSITION. *No structural causal model has a variable Y such that*

1. *Y depends on some variables*
2. *Dependence is dense at Y .*

PROOF. Suppose such a structural causal model existed. By (1), there is variable X with values x, x' and values o of the parents of Y other than X such that $f_Y(x, o) \neq f_Y(x', o)$. And by (2), there is a parent Z of Y such that $f_Y(x, z, o_{-z}) = f_Y(x', z, o_{-z})$, where o_{-z} are the values in o other than z . A contradiction follows:

$$f_Y(x, o) \stackrel{(1)}{\neq} f_Y(x', o) = f_Y(x', z, o_{-z}) \stackrel{(2)}{=} f_Y(x, z, o_{-z}) = f_Y(x, o).$$

□

Here are some case studies to help illustrate the impossibility. Let $y_{t'}$ be a spacetime point from the light example. The parents of $y_{t'}$ are the points x_t such that the light from x at t can reach y at t' (i.e. $t' = y - x/c$). Let's consider some candidate structural equation for y_t and see how they fail.

(\forall) $y_{t'} = 1$ just in case $x_t = 1$ for every parent x_t of $y_{t'}$.

Let x_t and $z_{t''}$ be parents of $y_{t'}$ with $t < t'' < t'$. Intuitively, intervening to make x dark at t and z lit at t'' makes y lit at t' : $[x_t = 0, z_{t''} = 1]y_{t'} = 1$. But the universal rule wrongly predicts that y is dark at t' under this intervention.

(\exists) $y_{t'} = 1$ just in case $x_t = 1$ for some parent x_t of $y_{t'}$.

Vice versa, intervening to make x lit at t and z dark at t'' intuitively results in y being dark at t' : $[x_t = 1, z_{t''} = 0]y_{t'} = 0$. But the existential rule wrongly predicts that $y_{t'}$ is lit at t' under this intervention.

($\exists\forall$) $y_{t'} = 1$ just in case there is a parent x_t of $y_{t'}$ such that $z_{t''} = 1$ for all parents $z_{t''}$ of $y_{t'}$ with $t < t'' < t'$.

Let x_t be a parent of $y_{t'}$ and consider a context where every spacetime point is dark. After intervening to make $x_t = 1$, the rule above is compatible with the assignment of values to variables where $x_t = 1$ and for every parent $z_{t''}$ of $y_{t'}$ with $t < t'' < t'$, $z_{t''} = 0$. In other words, the rule above allows that intervening to set $x_t = 0$ has no effect on any other variable.

($\forall\exists$) $y_{t'} = 1$ just in case for every parent x_t of $y_{t'}$ there is a parent $z_{t''}$ of $y_{t'}$ with $t < t'' < t'$ such that $z_{t''} = 1$.

This rule faces the same problem as the ($\exists\forall$) rule: intervening to set $x_t = 1$ is compatible with a context where $z_{t''} = 0$ for all parents $z_{t''}$ of $y_{t'}$ with $t < t'' < t'$, in which case the intervention on x_t has no effect on any later variable.

6.7.5 The impossibility of dense causal chains in Bayesian networks

We can prove an analogous impossibility result for Bayesian networks. Recall the definition of Bayesian networks (Definition 6.6.2): a Bayesian network is a triple (V, E, P) where (V, E) is a directed acyclic graph and P a joint probability distribution over V that Markov and minimal with respect to (V, E) : every variable is independent of its non-descendants conditional on its parents, and this no longer holds if we remove any edges from the graph.

Given this definition, it is easy to see why Bayesian networks cannot represent dense causal chains. Let us first compare the two graphs in Figure 6.44.

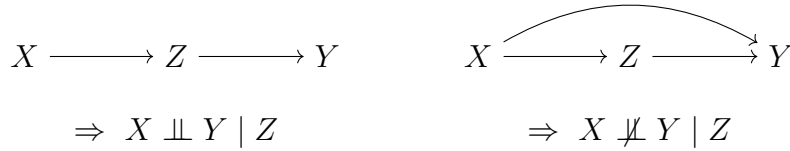


Figure 6.44

The Markov implies condition that in the graph on the left, conditioning on the middle variable screens off the variables on either side. Minimality implies that this does not hold in the graph on the right. Now, in a dense causal chain, dependence is always indirect: for any two variables, if X is a parent of Y , we can always find an intermediate variable Z that cuts off the dependence between X and Y . But then by minimality, X is not a parent of Y . Since the chain is dense, we can always find an intermediate variable between any variable and Y , so removing edges to satisfy minimality will lead to removing every edge into Y whatsoever, resulting in no dependence at all: if there are no edges into Y , the Markov condition predicts Y to be independent of every other variable.

In general, any probability distribution representing a dense causal chain (such as the example in Figure 6.39) will intuitively have the following property, which we call *dense dependence*.

6.7.3. DEFINITION. We define that a Bayesian network (V, E, P) has *dense dependence* just in case there is a variable Y such that (i) Y has a parent, and (ii) for any parent X of Y there is another parent Z of Y such that

1. X and Y are independent conditional on any set containing Z .
2. There is a directed path from X to Z .

A simple impossibility result follows.

6.7.4. PROPOSITION. *No Bayesian network has dense dependence.*

PROOF. Suppose there existed a Bayesian network (V, E, P) with dense dependence. Then there is some variable Y with parents X and Z and a directed path from X to Z .

Consider the graph $G^* = (V, E \setminus \{(X, Y)\})$ that results from removing the edge $X \rightarrow Y$ from G . G^* is a proper subgraph of G since $(X, Y) \in G$. We show that P is Markov with respect to G^* .

Note that every variable W other than Y has the same parents in G^* as in G . And since there is a directed path from X to Y , via $X \rightarrow \cdots \rightarrow Z \rightarrow Y$, every variable has the same non-descendants in G^* as in G . So to show that P is Markov with respect to G^* we only have to show that Y is independent of its non-descendants conditional on its parents in G^* . Since X and Y are independent conditional on any set containing Z , X and Y are independent conditional on the non-decedents and parents of Z . Pick any value y of Y , value nd_Y^* of Y 's non-descendants in G^* and value pa_Y^* of Y 's parents in G^* . Then

$$\begin{aligned}
 P(y \mid nd_Y^*, pa_Y^*) &= P(y \mid nd_Y, pa_Y^*) && (nd_Y^* = nd_Y) \\
 &= P(y \mid nd_Y, pa_Y^*, x) && (Z \in PA_Y^*, \text{ so } Y \perp\!\!\!\perp X \mid ND_Y, PA_Y^*) \\
 &= P(y \mid nd_Y, pa_Y) && (PA_Y = PA_Y^* \cup \{X\}) \\
 &= P(y \mid pa_Y) && (G \text{ is Markov relative to } P) \\
 &= P(y \mid pa_Y^*, x) && (PA_Y = PA_Y^* \cup \{X\}) \\
 &= P(y \mid pa_Y^*) && (Z \in PA_Y^*, \text{ so } Y \perp\!\!\!\perp X \mid PA_Y^*)
 \end{aligned}$$

Thus P is also Markov with respect to G^* . Since G^* is a proper subgraph of G , (V, E, P) violates minimality and is therefore not a Bayesian network. \square

6.7.6 Diagnosing the difference between structural causal models and the present framework

Why can our framework represent dense causal chains, while structural causal models cannot? The difference comes down to the following: in our framework, dependence emerges from the interplay of sudden changes (i.e. considering the A -variants) and nomic possibility, while structural causal models encode dependence directly via structural equations.

To appreciate the difference between encoding dependence directly and taking it to be emergent, it is instructive to consider the case of Thomson's lamp. Thomson (1954) imagines a being who can switch a lamp on and off with incredible speed. At time 0 the lamp is on. At the one-minute mark they switch it off (assume for simplicity that the lamp turns on instantaneously; if we do not wish to make this assumption we may simply add some delay between a change in the switch and a change in the lamp). Thirty seconds later they switch it on again, fifteen seconds after that they switch it off again, and so infinitely on. Thomson

then asks: after the being has completed infinitely many switches, i.e. at the two-minute mark, is the lamp on or off?

Thomson reasons as follows.

It seems impossible to answer this question. It cannot be on, because I did not ever turn it on without at once turning it off. It cannot be off, because I did in the first place turn it on, and thereafter I never turned it off without at once turning it on. But the lamp must be either on or off. This is a contradiction.

(Thomson 1954:5)

This, of course, is a mistake. As Benacerraf (1962:768ff.) points out, Thomson's reasoning only applies to times strictly before 2 minutes. Where $t_0 = 0$ and $t_1 = 2:00$, Benacerraf writes, "the only reasons Thomson gives for supposing that his lamp will not be off at t_1 are ones which hold only for times *before* t_1 Thomson's instructions do not cover the state of the lamp at t_1 , although they do tell us what will be its state at every instant *between* t_0 and t_1 (including t_0).". Benacerraf argues that Thomson's scenario is compatible with the lamp being on at 2:00 and with it being off at 2:00 – a point with which I agree.

To emphasise his point, Benacerraf imagines the following slightly different case. Suppose we wish to classify each number, including 0, as *foul* or *fair*, subject to the following constraint.

Consider the infinite converging sequence $1/2, 1/4, 1/8, 1/16, 1/32, \dots$. Its first member is foul, its second member fair, its third member foul, its fourth fair, etc., alternating in such a way that $1/2^n$ is foul if n is odd and fair if n is even, for all positive integers n . What about the limit of the sequence? It is, of course, not in the sequence; but is it foul or fair? ... The answer is simply that we haven't been told how to classify the limit number. The instructions cover the sequence and the sequence only. Nothing was said about any number not in the sequence. The same is true in the case of the lamp.

(Benacerraf 1962:769)

It is clear that determining whether each number in the sequence is foul or fair does not determine whether the limit of the sequence is foul or fair. What is surprising in the case of Thomson's lamp is that we naturally expect that the state of the lamp at any given time to be determined by the state of the world at previous times. This is nothing more than the assumption of determinism. This expectation is indeed met if we assume that the lamp can change only finitely many times within any interval $[0, t)$. For under that assumption, the lamp is on at time t just in case it was changed to on at the final time before t when the lamp was changed. However, in Thomson's scenario, 'the final time before t when the lamp was changed' does not exist. In everyday life scenarios it is

of course natural to assume that the lamp can change only finitely many times within any bounded interval. Thus the appearance of paradox in Thomson's lamp results from unwittingly extending an assumption that works in everyday life to supertasks. Allowing the lamp to change infinitely many times in an interval $[0, t)$ means that the state of the lamp at t is no longer determined by the state of the lamp at the previous times. In general, then, the lesson of Thomson's lamp is that supertasks can turn a deterministic system into a non-deterministic one.

With these observations it is easy to understand why structural causal models cannot represent dense causal chains. Let $C = \{x_t\}_{t \in [0,1]}$ be a set of variables representing the dense causal chain we wish to model. For example, in the light example above (Figure 6.39), each x_t represents whether point x in space is illuminated at time t . To represent the chain as a structural causal model, we need the value of each variable x_t , for $t > 0$, to be a function of the values of the previous variables, $PA_{x_t} = \{x_{t'}\}_{t' \in [0, t)}$. That is, we need a function $f_{x_t} : R(PA_{x_t}) \rightarrow R(x_t)$. Since functions must be defined for every input, f_{x_t} must also be defined under a supertask on the parents of x_t . To illustrate, consider the following assignment of values to the parents of x_t .

$$\begin{array}{lll} x_{t'} = 1 & \text{if} & 0 \leq t' < t/2 \\ x_{t'} = 0 & \text{if} & t/2 \leq t' < 3t/4 \\ x_{t'} = 1 & \text{if} & 3t/4 \leq t' < 7t/8 \\ x_{t'} = 0 & \text{if} & 7t/8 \leq t' < 15t/16 \\ & & \vdots \end{array}$$

That is, for any $t' \in [0, t)$ we let $x_{t'} = 1$ just in case $\max\{n \in \mathbb{N} : 1 - t/2^n \leq t'\}$ is even. Now, the function f_{x_t} must return a value for x_t given this assignment of values to its parents. But as we have seen from our discussion of Thomson's lamp, this assignment does not determine a value for x_t . Just as there is no way to determine whether Thomson's lamp is on or off at the completion of the supertask, there is no function that determines the value of x_t given the above assignment of values to its parents. Assuming Benacerraf's answer to Thomson's lamp (which I believe to be correct), then, we have reached a contradiction. A structural causal model representing a dense causal chain would have to, and yet cannot, determine the outcome of a Thomson's lamp-style supertask.

In summary, then, structural causal models cannot represent dense causal chains because doing so would require the value of a variable to be determined by *every* assignment of values to its parents – including those assignments that represent a supertask. But as we have seen, supertasks can leave the state of the system after their completion undetermined.

This also allows us to appreciate why we can represent dense causal chains in our framework. We mentioned above that in our framework, dependence emerges from the interplay of sudden changes (i.e. considering the A -variants) and nomic possibility. We do not encode dependence directly, as structural causal models do

via the structural equations. This means we can represent how the state of one point on a dense causal chain depends on a previous point without first having to determine how the point would behave under every possible combination of sudden changes, such as a supertask, i.e. after making sudden changes to infinitely many points prior to the given point. Unlike in structural causal models, in the present framework we can evaluate counterfactual assumptions on the fly as needed without first needing to specify the outcome of every counterfactual.

Indeed, our framework does not make any predictions about how a point in a dense causal chain would behave under a supertask on the previous points. For example, the framework does not make any predictions about the state of Thomson's lamp at 2:00 after making sudden changes at 1, 1:30, 1:45, and so on. Since such a series of changes intuitively does not determine the state of the lamp at 2:00, this is what we want.

Chapter 7

Exhaustification in the semantics of *cause* and *because*

Abstract. We show that a single operation can account for three seemingly distinct properties of the semantics of *cause* and *because*. The properties are, firstly, their comparative nature: interpreting *cause* and *because* involves comparing what would happen in the presence of the cause (a positive condition) with what would happen in the absence of the cause (a negative condition). Secondly, there is an asymmetry in logical strength between the two conditions: the positive condition involves a universal modal while the negative condition involves an existential modal. Thirdly, the positive and negative conditions have the same modal base, i.e. are interpreted while assuming the same set of background facts.

Despite their apparent dissimilarity, we show that these three properties are predicted by a single operation: exhaustification. The comparative nature of *cause* and *because* follows from the comparative nature of exhaustification, which compares a sentence with its alternatives. The asymmetry in strength arises because exhaustification negates alternatives: given the duality between universal and existential quantification, negation flips a necessity modal into a possibility modal, producing the observed strength asymmetry. Finally, the positive and negative condition have the same modal base since, rather *cause* and *because* having two modals in their semantics—one for the positive condition and one for the negative condition—their semantics contains a single modal which is copied by exhaustification.

We conclude by showing that this exhaustification operator violates Economy constraints, suggesting that it is not subject to licensing conditions but part of lexical semantics of *cause* and *because*.

7.1 Introduction

In this chapter we propose that three properties of the semantics of *cause* and *because* – properties that initially appear to have little to do with one another – are in fact the result of a single mechanism: an exhaustification operator in their lexical semantics. Let us introduce the three properties we aim to account for.¹

7.1.1 Three properties of *cause* and *because*

Property 1: The comparative nature of *cause* and *because*

A popular idea in the literature on causation is that the meaning of causal terms involves comparing what would happen in the presence of the cause versus what would happen in the absence of the cause. Recall the analyses we discussed in section 2.5.2, all of which have this shape, given in terms of a ‘positive condition’ and a ‘negative condition’.

Property 2: Asymmetry in strength between positive and negative conditions

The second property of the semantics of *cause* and *because* we consider is an asymmetry in strength between the positive condition and the negative condition. In section 2.7 we saw evidence that the positive condition is *strong* while the negative condition is *weak*, in the following sense. The positive condition requires that in *all* scenarios where the cause occurs the relevant condition is met, while the negative condition only requires that in *some* scenario where the cause does not occur the relevant condition is not met. What exactly this ‘relevant condition’ is depends on the analysis in question; for example, in the NESS and INUS conditions it is the effect occurring, according to Beckers it is that the cause produces the effect.

Property 3: The positive and negative conditions have the same background

Almost all analyses of causal claims appeal to some set of background facts (Suppes 1970, Cartwright 1979, Skyrms 1980, Mayrhofer et al. 2008). These facts are in some sense ‘taken for granted’ when evaluating a causal claim. To illustrate, consider (1).

- (1)
 - a. The light turned on because Alice flicked the switch.
 - b. Alice flicking the switch caused the light to turn on.

¹An extended version of this chapter has been published as McHugh (2023). Thanks to audiences at GLOW 44 and the Meaning Logic and Cognition seminar in Amsterdam for valuable feedback on the present chapter. A handout from the latter talk is available as McHugh (2021).

For these sentences to be true, one requires more than just the flicking of the switch. There must be power in the building, a wire connecting the switch and light, and so on. These background facts are involved in checking the positive condition. For example, in the NESS and INUS tests above one checks whether the presence of the cause is sufficient for the effect given some background facts: the ‘set’ in the words of the NESS test; the ‘condition’ in the words of the INUS test.

There is also a background involved when evaluating the negative condition: the facts from the actual world that are held fixed when evaluating what would happen if the cause had not occurred. For example, when interpreting (1) we consider relevant scenarios where Alice had not flicked the switch. In these scenarios the ‘background’ or ‘circumstances’ are held fixed; for instance, given that there is actually power in the building and a wire connecting the switch and light, one does not consider scenarios where there is no power in the building or no wire connecting the switch to the light.

In section 7.3 we see evidence that the positive condition is interpreted with respect to a background, and the negative condition is also interpreted with respect to a background. This raises the question whether there is any systematic relationship between the two. Section 7.3.2 presents evidence that these two backgrounds must be the same. One may stipulate that the lexical semantics of *cause* and *because* require them to be same, as for example the NESS and INUS tests do, where the ‘set’ or ‘condition’, minus the cause, is the same for both the positive and negative conditions. However, one may also wonder whether there is a more systematic principle accounting for the fact that the positive and negative backgrounds must be the same.

7.1.2 The apparent dissimilarity of properties 1, 2 and 3

These, then, are the three properties of *cause* and *because* we seek to account for. At first glance they are quite different. Granted, Property 1 is necessary for Properties 2 and 3, since Properties 2 and 3 are formulated in terms of the positive and negative conditions that are guaranteed by Property 1. But beyond this, the three properties appear to have nothing to do with one another. Property 2 is about the logical strength of the positive and negative conditions, while Property 3 is about their backgrounds. They appear to be about different aspects of the meaning of *cause* and *because*. For instance, one could imagine a lexical entry of *cause* and *because* that involves comparing the presence and absence of the cause (i.e. has Property 1), but does not require the positive and negative conditions to be evaluated with respect to the same set of background facts (lacks Property 3).

Similarly, one can imagine a lexical entry with Property 1 but without Property 2: such an entry would compare the presence and absence of the cause along some dimension (such as sufficiency for the effect) but would require both condi-

tions to be strong, or both conditions to be weak. Indeed, we have already seen an analysis of this kind: recall that Lewis (1973a) proposed that an event e causally depends on an event c just in case if c had occurred, e would have occurred, and if c had not occurred, e would not have occurred. In the same year that Lewis published his paper on causation (Lewis 1973a), he also published *Counterfactuals* (Lewis 1973b) in which he proposed that *would* is a necessity modal, requiring that in *all* the most similar worlds to the actual world where the antecedent holds, the consequent also holds. In the terminology above, Lewis’s analysis of causal dependence has both a strong positive condition and a strong negative condition. And the analysis does not say anything about Property 3. So Lewis’s analysis has Property 1 but lacks Property 2 and may or may not have Property 3. Lewis’ analysis of causal dependence does not violate any general theoretical principles. So it would be all the more surprising if the semantics of *cause* and *because* derived Properties 1, 2 and 3 from a single source.

This, however, is what we propose. We show that a single operator in the semantics of *cause* and *because* can account for all three properties.

7.1.3 Preliminaries: overview of the semantics of modality

Before presenting our proposed semantics of *cause* and *because*, let us briefly introduce the framework in which the semantics will be expressed. We will express our analysis in terms of Kratzer’s analysis of modality. We let $\Box_{f,g}(p)(q)$ be a universal counterfactual modal with modal base f , ordering source g , restrictor p and nuclear scope q . We interpret this claim as saying that p is sufficient for q according to our analysis of sufficiency from chapter 3. The reason for this notation is that it makes the modal’s parameters explicit, which will prove helpful in this chapter.²

7.2 *Cause, because, and exhaustification*

We will consider two semantics of *cause* and *because*, what we call the ‘simplified’ and the ‘full’ semantics. The simplified semantics is a useful first approximation of the meaning of *cause* and *because*, but faces well-known problems from overdetermination cases, discussed below. The full semantics overcomes these problems, and is our proposal for the meaning of *cause* and *because*. We begin with the simplified semantics.

²To simplify notation, we will sometimes omit the parameters and simply write $\Box(p)(q)$ for $\Box_{f,g}(p)(q)$.

7.2.1 The simplified semantics

On the simplified semantics, the positive condition states that the cause is sufficient for the effect given the background. We formalize this sufficiency condition as $\Box_{f,g}(p)(q)$.

Our key observation is that Properties 1, 2 and 3 above all follow from the presence of a single operator: exhaustification with respect to the cause's polar alternatives: $\{p, \neg p\}$. Exhaustification is defined as follows, and is akin to a silent *only*.³

$$O_{\text{ALT}} \varphi = \varphi \wedge \forall \psi \in \text{ALT}((\varphi \text{ does not entail } \psi) \rightarrow \neg \psi)$$

When we plug in the sufficiency condition $\Box_{f,g}(p)(q)$ for the prejacent and replace the cause p with its polar alternative $\neg p$, exhaustification checks whether the prejacent entails the result, $\Box_{f,g}(\neg p)(q)$: if not, exhaustification negates it.⁴

As it happens, $\Box_{f,g}(p)(q)$ does not entail $\Box_{f,g}(\neg p)(q)$, since it is possible for p to guarantee q given the circumstances while $\neg p$ does not also guarantee q given the circumstances. Thus exhaustification of the sufficiency condition $\Box_{f,g}(p)(q)$ has the following effect.⁵

$$O_{\{p, \neg p\}} \Box_{f,g}(p)(q) = \underbrace{\Box_{f,g}(p)(q)}_{\text{Simplified positive condition}} \wedge \underbrace{\neg \Box_{f,g}(\neg p)(q)}_{\text{Simplified negative condition}}$$

This condition states, loosely put, that given the circumstances, the cause is sufficient for the effect but the absence of the cause is not sufficient for the effect. This is essentially the NESS test above, formalized in terms of circumstantial modality.

Our simplified semantics for *because* and *cause*, given in (2), states that the cause occurred, and that this exhaustified sufficiency condition above holds. For brevity we only state the entries for *because*, though it should be understood as also applying to *cause* with the left and right arguments swapped (i.e. q *because* p

³For simplicity's sake we use Krifka's (1993) entry for *only*. Our results also follow from Fox's (2007) exhaustivity operator, based on the notion of innocent exclusion. For an overview and comparison of exhaustivity operators see Spector (2016).

⁴For brevity, we will write $O_{\{p, \neg p\}}\varphi$ where the alternatives are φ itself, and the result of substituting $\neg p$ for p in φ . We will also sometimes write $O_{\{p, \neg p\}}$ simply as O .

⁵The inference to the negative condition is reminiscent of *conditional perfection* (Geis and Zwicky 1971). For an approach to conditional perfection that uses exhaustification and assumes that *if not-p, q* is an alternative to *if p, q*, see Bassi and Bar-Lev (2018:§5). Note also that it does not matter whether we take the positive condition as the prejacent and derive the negative condition by exhaustification, or vice versa, take the negative condition as the prejacent and derive the positive condition by exhaustification: $O_{\{p, \neg p\}} \Box_{f,g}(p)(q)$ is equivalent to $O_{\{p, \neg p\}} \neg \Box_{f,g}(\neg p)(q)$.

is equivalent to p cause q).^{6,7}

(2) Semantics of *because* (simplified).

$$\llbracket \text{because} \rrbracket = \lambda p_{\langle s,t \rangle} \lambda q_{\langle s,t \rangle} : p \wedge O_{\{p, \neg p\}} \Box_{f,g}(p)(q).$$

7.2.2 Properties 1, 2, and 3 via exhaustification

Properties 1, 2 and 3 fall out immediately from exhaustification.

The comparative character of *because* (Property 1) results from the comparative nature of exhaustification, which compares the prejacent with its alternatives. We stipulate that in the semantics of *because*, the alternatives are the cause's polar alternatives.

⁶We place the condition that the cause occurred (p) outside the scope of exhaustification because otherwise exhaustification would be vacuous, as we see in the following chain of equivalences.

$$\begin{aligned} & O_{\{p, \neg p\}} (p \wedge \Box(p)(q)) \\ \Leftrightarrow & p \wedge \Box(p)(q) \wedge \neg(\neg p \wedge \Box(\neg p)(q)) \\ \Leftrightarrow & p \wedge \Box(p)(q) \wedge (p \vee \neg\Box(\neg p)(q)) \\ \Leftrightarrow & p \wedge \Box(p)(q) \end{aligned}$$

⁷Note that we do not need to add q as a conjunct to (2) since $p \wedge \Box_{f,g}(p)(q)$ entails q : if p is true and is sufficient for the truth of q , then q is also true.

Our entry for *because* in (2) assigns the same status to the condition that the cause occurred (p) as we do to the other conjunct $O_{\{p, \neg p\}} \Box_{f,g}(p)(q)$. Both are entailments. Alternatively, one might propose that p is encoded as a presupposition in the lexical semantics of *because*. Such a stipulation does not account for why some inferences rather than others are selected as presuppositions in the first place (see Abrusán 2011, 2016:for discussion). Moreover, the inferences from *cause* and *because* that their arguments are true is a soft presupposition in the sense of Abusch (2002, 2010), as they are easily suspendable, as shown in (i).

- (i) a. The outcry which followed *Morgan* was not because the House of Lords had changed the law but because the public mistakenly thought it had done so. (Source: Temkin 2002)
- b. No, the coronavirus did not cause the death rate to drop in Chicago... Overall, deaths don't appear to be declining. (Source: [Politifact.com](https://www.politifact.com), 3 April 2020)

Romoli (2012, 2015) proposes that the projection properties of *because* are in fact due to a scalar implicature. An utterance of $\neg(q \text{ because } p)$ triggers the alternatives $\neg p$ and $\neg q$. Since $\neg(q \text{ because } p)$ – whose meaning according to (2) is given in (iia) – entails neither alternative, we derive the implicatures in (iib).

- (ii) a. $\neg(q \text{ because } p) \Leftrightarrow \neg p \vee \neg\Box(p)(q) \vee \Box(\neg p)(q)$
- b. $O_{\text{ALT}} \neg(q \text{ because } p) \Leftrightarrow \neg(q \text{ because } p) \wedge p \wedge q$ where
 $\text{ALT} = \{\neg(q \text{ because } p), \neg p, \neg q\}.$

Given Romoli's account, we can capture the projection properties of *because* without needing to assign a special status to p in the lexical semantics of *because*.

The asymmetry in strength between the positive and negative conditions (Property 2) results from the fact that exhaustification negates alternatives. This parallels the behaviour of *only* when it composes with a universal modal, such as *guaranteed*:

- (3) a. You are guaranteed to get a seat only if you book in advance.
 (i) \Rightarrow You are **not** guaranteed to get a seat if you do not book in advance.
 (ii) \nRightarrow You are guaranteed to **not** get a seat if you do not book in advance.
 b. The effect is guaranteed to occur only if the cause occurs.
 (i) \Rightarrow The effect is **not** guaranteed to occur if the cause does not occur.
 (ii) \nRightarrow The effect is guaranteed to **not** occur if the cause does not occur.

In (3) we assume that broad focus on the *if*-clause triggers its polar alternative.⁸ Exhaustification, like *only*, negates the prejacent's excludable alternatives. Given the duality between universal and existential quantification, the negation contributed by exhaustification turns a necessity modal into a possibility modal, generating the observed asymmetry in strength: $O_{\{p, \neg p\}} \Box_{f,g}(p)(q)$ entails $\neg \Box_{f,g}(\neg p)(q)$ rather than $\Box_{f,g}(\neg p)(\neg q)$.

Finally, the fact that the positive and negative conditions have the same background (Property 3) falls out from the fact that exhaustification simply copies the modal's parameters – the modal base (*f*) and ordering source (*g*) – without altering them. If, as we argue in section 7.3.1 below, the background involved in the interpretation of *cause* and *because* is the modal base, then exhaustification ensures that this background is the same in both the positive and negative conditions.

7.2.3 The full semantics

The simplified semantics faces well-known problems from cases of overdetermination, where a causal claim is intuitively true even though the effect would still have occurred without the cause (for a discussion of overdetermination cases see chapter 2, section 2.4). For this reason we also consider a semantics designed to work in cases with and without overdetermination alike. For the full semantics, we borrow the notion of *production* from Beckers (2016) and Beckers and Vennekens (2018), inspired by Hall (2004). Beckers aims to analyze the truth conditions of *is a cause of* within the framework of structural causal models. What is most important to observe for present purposes is the overall shape of Beckers'

⁸For more on how polar alternatives are generated by interaction with focus, see e.g. Biezma and Rawlins (2012) and Kamali and Krifka (2020).

analysis, which consists of the following two conditions, here stated informally (for a formalization see Beckers and Vennekens 2018).

- (4) *p* is a cause of *q* just in case
- a. *p* produced *q*, and
 - b. If *p* had not occurred, $\neg p$ would not have produced *q*.

Beckers' key innovation is that (4) does not require that if the cause had not occurred, the effect would not have occurred; rather, it requires that if the cause had not occurred, the absence of the cause would not have produced the effect.⁹

At first glance, it is quite surprising that the semantics of *cause* and *because* would involve considering whether the absence of the cause would have itself produced the effect. Where could such a complex condition possibly come from? Notice that a formula of exactly this shape is expected if the semantics of *cause* and *because* involve considering whether the cause produced the effect, and also includes a mechanism that involves replacing the cause with its negation. Exhaustification with respect to the cause's polar alternatives is just such a mechanism.

However, clearly, exhaustifying (4a) does not result in (4b). The two conditions have a fundamentally different shape. Nonetheless, in section 2.2 we saw evidence that (4a) is not quite correct. The semantics of *cause* and *because* do not only require that the cause produce the effect, but requires that the truth of the cause be *sufficient* for the cause to produce the effect. Formally, the condition is $\Box_{f,g}(p)(p \text{ produce } q)$. When we exhaustify (4a), we do not get (4b), but when we exhaustify $\Box_{f,g}(p)(p \text{ produce } q)$, remarkably, we get exactly Beckers' condition in (4b).

$$O_{\{p, \neg p\}} \Box_{f,g}(p)(p \text{ produce } q) = \underbrace{\Box_{f,g}(p)(p \text{ produce } q)}_{\text{Full positive condition}} \wedge \underbrace{\neg \Box_{f,g}(\neg p)(\neg p \text{ produce } q)}_{\text{Full negative condition}}$$

If we replace *q* in the simplified semantics with *p produce q* we get the following semantic entry, which we call the 'full' semantics.

- (5) Semantics of *because* (full).
 $\llbracket \textbf{because} \rrbracket = \lambda p_{\langle s,t \rangle} \lambda q_{\langle s,t \rangle} : p \wedge O_{\{p, \neg p\}} \Box_{f,g}(p)(p \text{ produce } q).$

On the full semantics, Properties 1, 2 and 3 also fall out as a result of exhaustification, for the same reasons as on the simplified semantics.

7.2.4 Why put exhaustification in the semantics of *cause* and *because*?

What is to be gained by putting exhaustification into the semantics of *cause* and *because*? In one sense, quite little. The exhaustification operator is well-

⁹Sartorio (2005) and Weslake (2015) have previously proposed conditions similar to (4b).

defined, so one can always replace the exhaustified formula with its equivalent exhaustification-free result, if desired.

In another sense, however, writing the semantics of *cause* and *because* in terms of exhaustification allows us to derive some aspects of their meaning from a general mechanism, one not unique to causation or modality. Exhaustification appears in theories from a number of semantic domains, such as scalar implicatures (van Rooij and Schulz 2004, Schulz and van Rooij 2006, Spector 2007), polarity items (Krifka 1995, Chierchia 2013) and free choice inferences (Fox 2007). A number of authors go so far as to propose mandatory exhaustification, in the sense that every matrix sentence is parsed with an exhaustification operator by default (see Krifka 1995, Fox 2007, Magri 2009).

A compelling, albeit speculative idea is that the natural language system finds it economical to build meanings from familiar operations, with the more familiar the operation, the greater the gains in economy from its reuse. If we are constantly exhaustifying the sentences we interpret, as some have proposed, it is not so surprising to see the same operation appear in the lexical semantics of certain words. Exhaustification has previously been applied in the lexical semantics of Mandarin *dou* (Xiang 2016), approximative uses of *just* (Thomas and Deo 2020), and of course, *only*.¹⁰ If the present proposal is correct, we can add *cause* and *because* to the growing list of words whose meaning can be expressed in terms of exhaustification.

Of course, this kind of reasoning can only take us so far. Exhaustification alone does not tell us what to exhaustify, nor what the alternatives are.¹¹ Some motivation for polar alternatives – comparing the cause with its absence – may come from looking at the relationship between causal reasoning and decision-making. A paradigm case of causal reasoning concerns an agent deciding whether or not to do an action. Faced with a decision problem about whether or not to bring about *p*, we may think of the simplified semantics as addressing the questions *If I bring about p, will q be true? And if I do not, will q be true?* and the full semantics as addressing the questions *If I bring about p, will that produce q? And if I do not bring about p, will that produce q?*¹²

¹⁰More precisely, Thomas & Deo's entry for approximative *just* involves exhaustification in the sense that it exactly fits the definition of the exhaustification operator, when we take the alternatives to be levels of granularity rather than sentences. They propose that *just(p)* asserts that *p* is true at the finest level of granularity *g*, and for any granularity level *g'*, if *p* being true at *g* does not entail that *p* is true at *g'* then *p* is not true at granularity level *g'*. The parallel between this entry and exhaustification is striking.

¹¹Thanks to an anonymous *Glossa* reviewer for making this point.

¹²That being said, there are decision problems with more fine-grained alternatives; for instance, whether to take the train, tram, or metro, or whom to hire from a list of ten candidates (Thanks to Peter van Emde Boas for raising this issue.) Though we may think of polar alternatives as the least common denominator of all alternative sets, since we can always reframe a decision problem over a set of alternatives in terms of many decision problems, each with polar alternatives; for example, whether to take the train, tram or metro becomes whether or not to

Nonetheless, even if the paradigm case of causal reasoning involves polar alternatives, this still does not tell us why the paradigm case ends up hardwired into the semantics of *cause* and *because*. One may imagine an alternative meaning of *cause* and *because* which allows the set of alternatives to be contextually determined rather than fixed to be the cause's polar alternatives, $\{p, \neg p\}$. Since the definition of exhaustification allows for any set of alternatives, to derive the entries for *cause* and *because* we propose, we must add a stipulation that the alternatives used by exhaustification are the cause's polar alternatives. It remains to be seen whether this stipulation can be derived from general principles.

7.2.5 Comparing the full and simplified semantics

Before moving on to the data, let us pause to better understand the relationship between the full and simplified semantics.

Where the simplified semantics cares about whether or not the effect occurred, the full semantics cares about whether or not the cause produced the effect. It turns out that the two semantic entries are logically independent, in the sense that there are cases where the full semantics is satisfied but not the simplified semantics, and vice versa. Figure 7.1 shows entailment relations between the conditions of the full and simplified semantics. These are guaranteed by the two facts in (6).

- (6) a. **Production is factive:** p produce q entails $p \wedge q$.
 b. **Modals are upward entailing in their scope:**
 If q^+ entails q then $\Box(p)(q^+)$ entails $\Box(p)(q)$.

	Positive condition	Negative condition
<i>Full semantics</i>	$\Box(p)(p \text{ produce } q)$	$\neg\Box(\neg p)(\neg p \text{ produce } q)$
	\Downarrow	\Uparrow
<i>Simplified semantics</i>	$\Box(p)(q)$	$\neg\Box(\neg p)(q)$

Figure 7.1: Entailment relations between the parts of the full and simplified semantics.

It follows that the full semantics has a stronger positive condition but a weaker negative condition compared with the simplified semantics, as shown in Figure 7.1.

A further property of both semantics is that counterfactual dependence, here formalized as $\Box(\neg p)(\neg q)$, entails the negative conditions of both the full and simplified semantics. Let us start by showing that counterfactual dependence

take the train, whether or not to take the tram, and whether or not to take the metro.

entails the simplified negative condition, $\neg\Box(\neg p)(q)$. It is commonly assumed that modals – like all quantificational elements – presuppose that their domain is nonempty (Cooper 1983, von Fintel 1994, Beaver 1995, Ippolito 2006). In this case, the domain of the modal at world w is the set of worlds selected by the modal when restricted by the proposition p . Let us denote this set of worlds by $D(p, w)$. Then $\Box(\neg p)(\neg q)$ is true just in case *all* worlds in $D(\neg p, w)$ are $\neg q$ -worlds. We assume that an utterance that entails $\Box(\neg p)(\neg q)$ presupposes that $D(\neg p, w)$ is nonempty. So counterfactual dependence, $\Box(\neg p)(\neg q)$, together with the nonempty domain presupposition implies that *some* world in $D(\neg p, w)$ is a $\neg q$ -world. Since worlds are logically consistent, this world is not a q -world. So it is not the case that every world in $D(\neg p, w)$ is a q -world: $\neg\Box(\neg p)(q)$, which is just the simplified negative condition.

To see that counterfactual dependence entails the full negative condition (when the restricted modal has nonempty a domain), first recall that production is factive (6a). Contrapositively, if q does not occur then nothing produces q to occur; in particular, $\neg p$ does not produce q to occur. Thus $\neg q$ entails $\neg(\neg p \text{ produce } q)$. Then as modals are upward entailing in their scope (6b), we have the following chain of implications.

$$\begin{aligned}
 & \Box(\neg p)(\neg q) \\
 \Rightarrow & \Box(\neg p)(\neg(\neg p \text{ produce } q)) && \text{Modals are upward entailing in their scope} \\
 \Rightarrow & \neg\Box(\neg p)(\neg p \text{ produce } q) && \text{Nonempty domain assumption}
 \end{aligned}$$

The last formula is the full negative condition.

7.3 The positive and negative conditions have the same background

In our description of sufficiency in section 2.2, we stated that p is sufficient for q in a circumstance just in case, in that circumstance, it is not nomically possible for p to be true without q being true. In this section we analyze what “the circumstances” are. It is important to understand what determines the circumstances when we discuss Property 3: that the circumstances (or ‘background’) of the positive and negative conditions are the same.

The robot scenario from section 2.2 illustrates why we need to relativize nomic possibility to the circumstances. There we considered two contexts: one where the robot turns at random, and one where it always changes direction. We can represent these two contexts in two separate worlds. Then what is nomically possible in one world is nomically impossible in the other world (e.g. the robot taking Road A). We can capture this fact since nomic possibility is relative to a world, which is something already built into Kratzer’s (1981) account of modality.

However, we can also represent the two robot contexts in the same world. Suppose that one day the robot is programmed to turn at random, and the next day it is reprogrammed to always change direction. Then something that was nomically possible in one world at one time is no longer nomically possible in the same world at another time (e.g. the robot taking Road A). This shows that nomic possibility is relative to more than just the world of evaluation. It is also relative to what we may call *the circumstances*.

We would like to understand what determines the circumstances. To that end, consider the following scenario. For the sake of continuity, we will adapt the robot case. Suppose Roads B and D are lined with trees (see Figure 7.2). When the robot must choose whether to take a road with trees or one without, it is programmed to always take the road with trees. Otherwise it decides at random. On Monday it was positioned at the starting point. On Tuesday it took one of the roads in front of it. Since it faced two bare roads, it turned at random. On this particular occasion it happened to turn left. On Wednesday it faced Roads A and B: a bare road and one with trees. On Thursday it took one of the roads. Given the robot's programming, it took the road with trees, Road B. Consider (7) and (8) in this context.

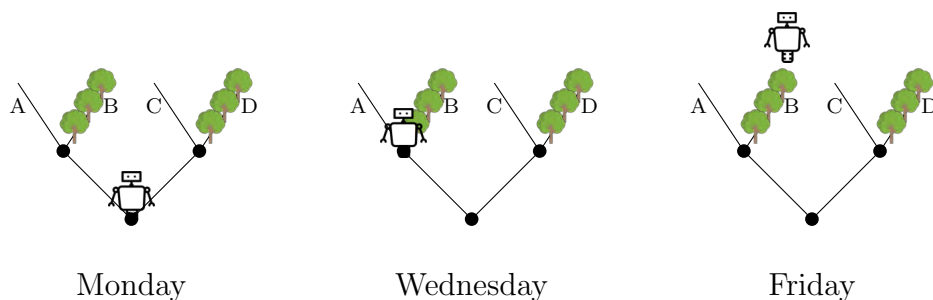


Figure 7.2: Context 1.

- (7)
 - a. Given how things were on Monday, the robot took Road B because it is programmed to prefer tree-lined roads.
 - b. Given how things were on Wednesday, the robot took Road B because it is programmed to prefer tree-lined roads.
- (8)
 - a. Given how things were on Monday, the robot's preference for tree-lined roads caused it to take Road B.
 - b. Given how things were on Wednesday, the robot's preference for tree-lined roads caused it to take Road B.

There is a contrast between the (a) sentences and the (b) sentences. Intuitively, the (a) sentences are false and the (b) sentences are true.

Intuitively, the (a) sentences are false because given how things were on Monday, the robot could have turned right, in which case it would have faced Roads

C and D, and so wouldn't have taken Road B. That is, the (a) sentences are false because they violate sufficiency.

The (b) sentences do not violate sufficiency because, given how things were on Wednesday, the robot was facing Roads A and B, so its programming guaranteed that it take Road B. Assuming that the (b) sentences satisfy the other requirements of *cause* and *because*, they are predicted to be true.

Given our analysis of sufficiency, the (a) sentences violate sufficiency because the fact that the robot turned left on Tuesday is not part of the circumstances used to interpret the (a) sentences, but is part of the circumstances used to interpret the (b) sentences. We can account for this difference by proposing that the *given*-clause determines the circumstances in each case. The circumstances for the (a) sentences are how things were on Monday, and the circumstances for the (b) sentences are how things were on Wednesday.

Observe that we can specify the circumstances with temporal information alone (e.g. *on Monday*, *on Wednesday*). This shows that the circumstances are a function of time. Of course, time is not enough: temporal information by itself (e.g. that it is Tuesday) does not carry any information unless one knows what world we are talking about. We see this in (7) and (8): the expression *how things were on Monday/Wednesday* refers to how things were on Monday/Wednesday in the world of evaluation.

7.3.1 The circumstances as modal base

This observation that the circumstances are a function of time is expected if what we have been calling 'the circumstances' are just the modal base of the modal expressed by *cause* and *because*. For it is often assumed that modals bases are sensitive to time. To see this, let us briefly review Condoravdi's account of the interaction between tense and modality.

Condoravdi (2002) proposes that modals have a temporal perspective and a temporal orientation. The temporal perspective is the time when the possibilities are evaluated. The temporal orientation is the relationship between the temporal perspective and the time of the embedded eventuality. Consider (9).

- (9) He might have won the game. (Condoravdi 2002:ex. 6)

Condoravdi (2002:62) points out that (9) has two readings, which she calls 'epistemic' and 'counterfactual'. On the epistemic reading, (9) describes the speaker's present knowledge about a past event: *might* has a present perspective and a past orientation. On the counterfactual reading, (9) describes what was possible at some point in the past – e.g. at half-time in the game – about an event in the future of that point: *might* has a past perspective and a future orientation. The two readings can be brought out with the following continuations (Condoravdi 2002:ex. (7)).

- (10) a. He might have (already) won the game (# but he didn't).
Epistemic reading: present perspective, past orientation
 b. At that point he might (still) have won the game but he didn't in the end.
Counterfactual reading: past perspective, future orientation

Following Condoravdi (2002:71), we assume that modal bases are a function of the world of evaluation and the temporal perspective. We can then capture the fact that the circumstances are a function of time if what we have been calling ‘the circumstances’ are the modal base of the modal expressed by *because*.

If we assume that the *given*-clauses in (7) and (8) set the modal base, we can account for the contrast between the (a) and (b) sentences. The proposition that the robot turned left on Tuesday is not part of how things were on Monday, but is part of how things were on Wednesday.

The fact that the *given*-clauses can manipulate the modal base shows that the embedded causal claims, given in (11), are compatible with multiple modal bases.

- (11) a. The robot took Road B because it is programmed to prefer tree-lined roads.
b. The robot's preference for tree-lined roads caused it to take Road B.

Suppose that we are evaluating (11) after the robot has completed its journey. The sentences in (11), without the *given*-clause, do not specify when the possibilities are to be evaluated. That is, they do not specify the modal’s temporal perspective. If it is set to before the robot turned left on Tuesday, our proposed semantics for *because* (both the full and simplified versions) predict (11) to be false. If they are evaluated after the robot turns left on Tuesday, our proposed semantics predicts them to be true. The prediction that (11) are ambiguous appears to be correct. We can bring out the two readings as follows.

- (12) a. The robot took Road B because it is programmed to prefer tree-lined roads. For, its programming made it take Road B rather than Road A.
- b. The robot didn't take Road B because it is programmed to prefer tree-lined roads. For it could have turned right on Tuesday, in which case it would have taken Road C or D, not Road B.

We can also show it is possible to set the temporal perspective to Monday by modifying the scenario. Let us remove Road A and add trees to First Street, as in Figure 7.3.

Consider (11), repeated below, in this context.

- (11) a. The robot took Road B because it is programmed to prefer tree-lined roads.

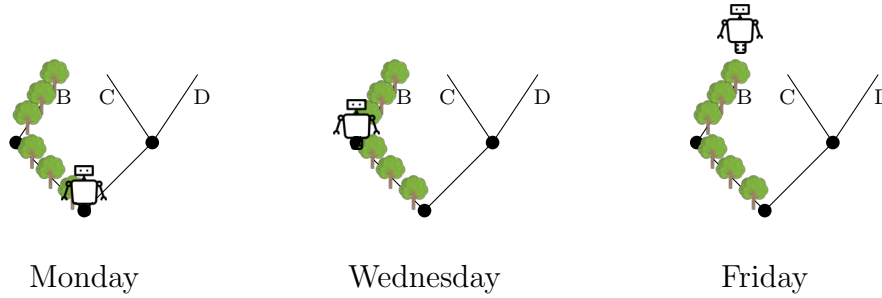


Figure 7.3: Context 2.

- b. The robot's preference for tree-lined roads caused it to take Road B.

In this scenario (11) has a true reading.

If the temporal perspective is set to Monday, the full and simplified semantics correctly predict this result. For then the positive conditions are satisfied, since the robot's programming guarantees that it take Road B. And the negative conditions are satisfied, since if the robot had not been programmed to prefer tree-lined roads, it could have taken Roads C or D. However, if the temporal perspective is set to Wednesday, the negative conditions are not satisfied. Given how things were on Wednesday, even if the robot had not been programmed to prefer tree-lined roads, Road B was its only option.

This provides further evidence that *because* and *cause* do not fix the temporal perspective of their modals.

7.3.2 Testing whether the positive and negative backgrounds can differ

Let us consider one last modification of the robot scenario. Suppose that Road A is still removed, and this time there are trees only on Road B, as depicted in Figure 7.4. As before, the robot is programmed to prefer tree-lined roads.

On Monday the robot first faced two bare roads. On this particular day it turned left, though it could just as easily have turned right. Then on Wednesday the robot faced a single tree-lined road, Road B, so it took it. At that point Road B was its only choice, so even if it hadn't preferred tree-lined roads, it would still have taken Road B.

Consider (11) in this context.

- (11) a. The robot took Road B because it is programmed to prefer tree-lined roads.
 b. The robot's preference for tree-lined roads caused it to take Road B.

Intuitively, (11) are false in this context.

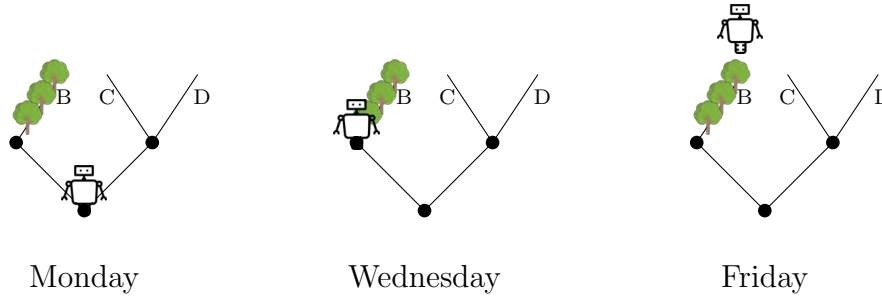


Figure 7.4: Context 3.

Let us consider what our semantics above has to say about this case. (Since there is no overdetermination in this scenario, the full and simplified semantics give the same verdict.) We saw above that when interpreting (11), it is possible to set the temporal perspective to Monday, before the robot made its first turn, and it is possible to set it to Wednesday, after the robot made its first turn.

Suppose the temporal perspective is set to Monday. Then the positive condition is false, since the robot could have turned right first and avoided Road B. But the negative condition is true, since if the robot hadn't been programmed to prefer tree-lined street, it could have avoided Road B by taking Roads C or D.

Suppose instead that the temporal perspective is set to Wednesday. Now the situation is reversed. The positive condition is true, since on Wednesday the robot was guaranteed to take Road B. But the negative condition is false, since if the robot hadn't been programmed to prefer the tree-lined street, it would still have taken Road B. These results are summarized in Table 7.1.

Temporal perspective	Positive condition	Negative condition
Monday	✗	✓
Wednesday	✓	✗

Table 7.1: For each temporal perspective, either the positive or negative condition fails.

If the temporal perspectives of the positive and negative conditions could differ when interpreting *cause* or *because*, we would expect (11) to have a true reading in the context of Figure 7.4.

Let us consider a second scenario, to help ensure that the arguments above are not due to particularities of the robot context. Suppose there are two veterinary clinics in Alice's region, one in village A and one in village B, each with two kinds of positions, junior and senior. (For simplicity, suppose that these four jobs are the only jobs Alice could have; for example, if Alice hadn't been a senior vet she would have been a junior vet.) The annual salaries for each position are listed in Table 7.2.

	Village A	Village B
Senior vet	30,000	20,000
Junior vet	15,000	15,000

Table 7.2: Salaries in context 1.

Actually, Alice works in village A and is a senior vet. Consider (13).

(13) Alice earns 30,000 per year because she's a senior vet.

Given the salaries in Table 7.2, (13) has a true reading.

(13) satisfies the negative condition: if Alice hadn't been a senior vet she would have earned 15,000 per year instead. For (13) to satisfy the positive condition the modal base must include the fact that Alice works in village A. If we considered the possibility of Alice working as a senior vet at village B she would earn 20,000, not 30,000. This is further illustrated by the fact that (14) is intuitively true.

(14) Given that Alice works in village A, she earns 30,000 per year because she's a senior vet.

Now consider (13) with respect to the salaries in Table 7.3.

	Village A	Village B
Senior vet	30,000	30,000
Junior vet	30,000	15,000

Table 7.3: Salaries in context 2.

Given the salaries in Table 7.3, again (13) has a true reading. This time the situation is reversed: (13) satisfies the positive condition regardless whether we fix the fact that Alice works in village A. But for (13) to satisfy the negative condition, we must not fix the fact that she works in village A. This is illustrated by the fact that (14) is intuitively false given the salaries in Table 7.3.

To summarize, the fact that (13) has a true reading for both salary tables above shows that the modal base of the modal in *because* is flexible: it can include or omit the fact that Alice works in village A.

Now consider (13) with respect to the salaries in Table 7.4.

	Village A	Village B
Senior vet	30,000	20,000
Junior vet	30,000	15,000

Table 7.4: Salaries in context 3.

Intuitively, (13) is false in this context.

Suppose the fact that Alice works in village A is part of the modal base when interpreting (13). Then the positive condition holds: given that she works in village A, the fact that she is a senior vet guarantees that she earns 30,000 per year. But the negative condition fails: if she were not a senior vet, she would be a junior vet (assuming these are the only positions available) and would still earn 30,000.

Suppose instead that the fact that Alice works in village A is not part of the modal base when interpreting (13). Now the situation is reversed. The positive condition fails since Alice could have been a senior vet in village B and would have only earned 20,000 per year. But the negative condition holds, since if Alice hadn't been a senior vet, she could have been a junior vet in village B and would have earned 15,000 per year.

These two reasons for (13)'s falsity are illustrated by the following continuations, uttered in a situation where the salaries are given by Table 7.4.

- (15) A: Alice earns 30,000 per year because she's a senior vet.
 a. B: That's not right. Even if she worked as a junior vet, she would still earn 30,000 per year.
 b. B': That's not right. The senior vets in village B only earn 20,000 per year.

To summarize, the fact that (13) has a true reading with respect to the salaries in Tables 7.2 and 7.3, provides evidence that the modal bases of the positive and negative conditions are flexible: they may include or omit the fact that Alice works in village A. But whichever it is, the modal's parameters in the positive and negative condition must be the same.

The fact that the modals of the positive and negative condition must have the same modal base is exactly what we expect from exhaustification. Even though the semantics of *cause* and *because* involves two modals (one in the positive condition and one in the negative condition), using exhaustification we may propose that their semantics in fact contains a single modal, which is copied by exhaustification. Since exhaustification only modifies the cause – replacing p with $\neg p$ – it copies the modal without touching its parameters f and g .

- (16) Semantics of *because* (simplified).

$$\llbracket \mathbf{q \text{ because } p} \rrbracket = p \wedge O_{\{p, \neg p\}} \Box_{f,g}(p)(q)$$

$$= p \wedge \Box_{f,g}(p)(q) \wedge \neg \Box_{f,g}(\neg p)(q)$$

- (17) Semantics of *because* (full).

$$\llbracket \mathbf{q \text{ because } p} \rrbracket = p \wedge O_{\{p, \neg p\}} \Box_{f,g}(p)(p \text{ produce } q)$$

$$= p \wedge \Box_{f,g}(p)(p \text{ produce } q) \wedge \neg \Box_{f,g}(\neg p)(\neg p \text{ produce } q)$$

If the semantics of *because* does not include exhaustification, then each modal is generated independently, as in (18).

- (18) a. $p \wedge \Box_{f,g}(p)(q) \wedge \neg \Box_{f',g'}(\neg p)(q)$
 b. $p \wedge \Box_{f,g}(p)(p \text{ produce } q) \wedge \neg \Box_{f',g'}(\neg p)(\neg p \text{ produce } q)$

Since each modal comes with a modal base and ordering source, without further constraints is conceivable that the two modals could differ in their parameters. To avoid this possibility we may, of course, add a constraint that the modals' parameters must be identical as a stipulation.

- (19) a. $p \wedge \Box_{f,g}(p)(q) \wedge \neg \Box_{f',g'}(\neg p)(q) \wedge \underline{f = f' \wedge g = g'}$
 b. $p \wedge \Box_{f,g}(p)(p \text{ produce } q) \wedge \neg \Box_{f',g'}(\neg p)(\neg p \text{ produce } q) \wedge \underline{f = f' \wedge g = g'}$

Now, it is reasonable to expect that two modals within the same lexical entry would be subject to such a constraint, forcing their parameters to be the same. The benefit of writing the semantics of *cause* and *because* using exhaustification is that we derive this constraint automatically, without needing to add it as a separate requirement.¹³

7.4 Economy

The previous sections provided evidence that the semantics of *cause* and *because* satisfies properties 1, 2 and 3. These three properties all point to the presence of an exhaustification operator in the lexical semantics of *cause* and *because*. One may wonder about the status of this operator. It is always present, or subject to licensing conditions?

To answer this question, a key test case is how *cause* and *because* behave under negation. It is commonly assumed that exhaustification is subject to an economy condition that prevents it from appearing when it would lead to an overall weaker meaning (Chierchia 2013, Fox and Spector 2018). If the exhaustification operator in the semantics of *cause* and *because* is subject to this constraint, we would expect the following parses of *cause* and *because* under negation to be ruled out by Economy.

$$\neg(p \wedge O_{\{p, \neg p\}} \Box(p)(q)) = \neg p \vee \neg \Box(p)(q) \vee \underline{\Box(\neg p)(q)}$$

$$\neg(p \wedge O_{\{p, \neg p\}} \Box(p)(p \text{ produce } q)) = \neg p \vee \neg \Box(p)(p \text{ produce } q) \vee \underline{\Box(\neg p)(p \text{ produce } q)}$$

Without exhaustification the underlined disjunct disappears:

$$\neg(p \wedge \Box(p)(q)) = \neg p \vee \neg \Box(p)(q)$$

$$\neg(p \wedge \Box(p)(p \text{ produce } q)) = \neg p \vee \neg \Box(p)(p \text{ produce } q)$$

Under negation, then, exhaustification in the semantics of *cause* and *because* leads to a weaker meaning.

However, it turns out that the only parse of *not ... because* and *not ... cause* that correctly accounts for the data is one that violates Economy, as we see now.

¹³Thanks to an anonymous *Glossa* reviewer for helpful discussion on this point.

7.4.1 *Because* and economy: data

In de Saint-Exupéry's *The Little Prince*, the protagonist visits a king who claims to be able to command the sun to set. Suppose the king commands the sun to set, and sure enough, some time later it sets. Unfortunately for the king's ego, the following sentences are false.

- (20) a. The sun set because the king commanded it.
 b. The king's command caused the sun to set.

The simplified and full semantics account for the falsity of (20) in different ways. On the simplified semantics (20) are false since the sun would have set even if the king hadn't commanded it; in symbols, $\Box(\neg command)(sunset)$. The simplified negative condition fails:

$$\underbrace{\Box(command)(sunset)}_{\text{Simplified positive condition: } \checkmark} \wedge \underbrace{\neg\Box(\neg command)(sunset)}_{\text{Simplified negative condition: } \times}$$

While on the full semantics (20) are false because the king's command did not produce the sun to set. This implies that the king's command is not sufficient for it to produce the sun to set; in symbols, $\neg\Box(command)(command \text{ produce } sunset)$.¹⁴ The full positive condition fails:

$$\underbrace{\Box(command)(command \text{ produce } sunset)}_{\text{Full positive condition: } \times} \wedge \underbrace{\neg\Box(\neg command)(\neg(command \text{ produce } sunset))}_{\text{Full negative condition: } \checkmark}$$

Compare this with the train track scenario. According to the simplified semantics, (20) are false for the same reason that (4), repeated below, are false in the train track scenario: the train would have reached the station anyway.

- (4) a. The train reached the station because the engineer flipped the switch.
 b. The engineer flipping the switch caused the train to reach the station.

While according to the full semantics, (4) and (20) are false for a different reason. Pulling the lever produced the train to reach the station (because there is a chain of events beginning with the engineer pulling the lever, through the train taking the side track, to the train reaching the station). But symmetrically, not pulling the lever would have also produced the train to reach the station, so the full

¹⁴This follows from modus ponens for universal modals, (20), together with the fact that the king did command the sun to set: $command \wedge \neg(command \text{ produce } sunset)$ entails $\neg\Box(command)(command \text{ produce } sunset)$.

semantics predicts (4) to be false.

$$\begin{array}{ccc}
 \underbrace{\Box(\textit{pull})(\textit{reach station})}_{\text{Simplified positive condition: } \checkmark} & \wedge & \underbrace{\neg\Box(\neg\textit{pull})(\textit{reach station})}_{\text{Simplified negative condition: } \times} \\
 \\
 \underbrace{\Box(\textit{pull})(\textit{pull produce reach station})}_{\text{Full positive condition: } \checkmark} & \wedge & \underbrace{\neg\Box(\neg\textit{pull})(\neg(\textit{pull produce reach station}))}_{\text{Full negative condition: } \times}
 \end{array}$$

Putting these sentences under negation, we observe that the following sentences are intuitively true (where *not ... because* is read with *not* scoping above *because*).

- (21) a. The sun did not set because the king commanded it.
 b. The king's command did not cause the sun to set.
- (22) a. The train did not reach the station because the engineer flipped the switch.
 b. The engineer flipping the switch did not cause the train to reach the station.

With these data at hand, let us see which parses using exhaustification account for them.

7.4.2 *Because* and economy: analysis

For the simplified semantics, the above data are compatible with two parses. The first, $\neg(p \wedge O\Box(p)(q))$, violates Economy.¹⁵ The second, which Fox and Spector (2018:ex. 70) discuss, features a higher exhaustification operator whose alternative is the prejacent without exhaustification: $O_{ALT}\neg O\Box(p)(q)$, where $ALT = \{\neg O\Box(p)(q), \neg\Box(p)(q)\}$. In essence, the higher operator adds that the lower operator was required for the sentence to be true. This parse does not violate Economy.

Table 7.5 gives four possible parses of *not ... because* with exhaustification, what truth value each predicts for (21) and (22), and whether the parse satisfies Economy.

Parse	Simplified meaning	(21)	(22)	Economy
$\neg\Box(p)(q)$	$\neg\Box(p)(q)$	F \times	T \checkmark	\checkmark
$O\neg\Box(p)(q)$	$\neg\Box(p)(q) \wedge \Box(\neg p)(q)$	F \times	F \times	\checkmark
$\neg O\Box(p)(q)$	$\neg\Box(p)(q) \vee \Box(\neg p)(q)$	T \checkmark	T \checkmark	\times
$O_{ALT}\neg O\Box(p)(q)$	$\Box(p)(q) \wedge \Box(\neg p)(q)$	T \checkmark	T \checkmark	\checkmark

Table 7.5: Possible parses of *not ... (be)cause* on the simplified semantics.

¹⁵To avoid notational clutter we write $O_{\{p, \neg p\}}$ simply as O .

The table shows that two parses of *not ... because* on the simplified semantics correctly predict (22) and (21) to be true, $\neg O\Box(p)(q)$ and $O_{ALT}\neg O\Box(p)(q)$, with the latter satisfying Economy.

This changes when we turn to the full semantics. Table 7.6 shows that only one parse of the full semantics correctly predicts the truth of (21) and (22). This is also the only parse that violates Economy. In the table we use $p \rightsquigarrow q$ as shorthand for p produce q . As above, we consider the parse $O_{ALT}\neg O\Box(p)(p \rightsquigarrow q)$ where $ALT = \{\neg O\Box(p)(p \rightsquigarrow q), \neg\Box(p)(p \rightsquigarrow q)\}$.

Parse	Full meaning	(21)	(22)	Economy
$\neg\Box(p)(p \rightsquigarrow q)$	$\neg\Box(p)(p \rightsquigarrow q)$	T ✓	F ✗	✓
$O\neg\Box(p)(p \rightsquigarrow q)$	$\neg\Box(p)(p \rightsquigarrow q) \wedge \Box(\neg p)(p \rightsquigarrow q)$	F ✗	F ✗	✓
$\neg O\Box(p)(p \rightsquigarrow q)$	$\neg\Box(p)(p \rightsquigarrow q) \vee \Box(\neg p)(p \rightsquigarrow q)$	T ✓	T ✓	✗
$O_{ALT}\neg O\Box(p)(p \rightsquigarrow q)$	$\Box(p)(p \rightsquigarrow q) \wedge \Box(\neg p)(p \rightsquigarrow q)$	F ✗	T ✓	✓

Table 7.6: Possible parses of *not ... (be)cause* on the full semantics.

We saw in section 2.4 that the full semantics is superior to the simplified semantics in overdetermination cases (i.e. cases without counterfactual dependence where the causal claim is nonetheless true). Assuming, then, that the full semantics is the correct semantics of *cause* and *because*, Table 7.6 shows that the exhaustification operator in the semantics of *cause* and *because* violates Economy. This is not so surprising if exhaustification is hard-coded into the lexical semantics of these words, making it obligatory even when it leads to an overall weaker meaning.

7.5 Conclusion

In this chapter we presented an account of three properties of *cause* and *because*.

Property 1. The comparative nature of *cause* and *because*.

The semantics of *cause* and *because* involves comparing what would happen in the presence of the cause with what would happen in its absence.

Property 2. The asymmetry in strength between the two conditions.

The positive condition has universal modal force while the negative condition has existential modal force.

Property 3. The positive and negative conditions have the same background.

The facts from the actual world that are held fixed when evaluating the positive and negative conditions are the same.

On the surface, these properties are all quite different. However, we saw that we can account for them all in a uniform way, by proposing that there is an exhaustification operator in the lexical semantics of *cause* and *because*. Now, we are not forced to write their semantics of these words in terms of exhaustification; as discussed in section 7.2.4, we can always rewrite their semantics without exhaustification if desired. But doing so allows us to account for three features of their semantics using a domain-general operation, one not unique to causality or modality. In a sense, then, what we have shown is that the meanings of *cause* and *because* are more ordinary than we may have imagined.

Bibliography

- Abrusán, Márta (2011). Predicting the presuppositions of soft triggers. *Linguistics and philosophy* 34.6, pp. 491–535. DOI: [10.1007/s10988-012-9108-y](https://doi.org/10.1007/s10988-012-9108-y).
- (2016). Presupposition cancellation: explaining the ‘soft–hard’ trigger distinction. *Natural Language Semantics* 24.2, pp. 165–202. DOI: [10.1007/s11050-016-9122-7](https://doi.org/10.1007/s11050-016-9122-7).
- Abusch, Dorit (1997). Sequence of tense and temporal de re. *Linguistics and philosophy*, pp. 1–50. DOI: [10.1023/A:1005331423820](https://doi.org/10.1023/A:1005331423820).
- (1998). Generalizing tense semantics for future contexts. *Events and grammar*. Ed. by Susan Rothstein. Springer, pp. 13–33. DOI: [10.1007/978-94-011-3969-4_2](https://doi.org/10.1007/978-94-011-3969-4_2).
- (2002). Lexical alternatives as a source of pragmatic presuppositions. *Semantics and linguistic theory*. Vol. 12, pp. 1–19. DOI: [10.3765/salt.v12i0.2867](https://doi.org/10.3765/salt.v12i0.2867).
- (2010). Presupposition Triggering from Alternatives. *Journal of Semantics* 27.1, pp. 37–80. DOI: [10.1093/jos/ffp009](https://doi.org/10.1093/jos/ffp009).
- Alonso-Ovalle, Luis (2006). Disjunction in alternative semantics. PhD thesis. University of Massachusetts Amherst. URL: <http://people.linguistics.mcgill.ca/~luis.alonso-ovalle/papers/alonso-ovalle-diss.pdf>.
- (2009). Counterfactuals, correlatives, and disjunction. *Linguistics and Philosophy* 32.2, pp. 207–244. DOI: [10.1007/s10988-009-9059-0](https://doi.org/10.1007/s10988-009-9059-0).
- Andreas, Holger and Mario Günther (2020). Causation in terms of production. *Philosophical Studies* 177.6, pp. 1565–1591. DOI: [10.1007/s11098-019-01275-3](https://doi.org/10.1007/s11098-019-01275-3).
- (2021). Difference-Making Causation. *The Journal of Philosophy* 118 (12), pp. 680–701. DOI: [10.5840/jphil20211181243](https://doi.org/10.5840/jphil20211181243).
- Anscombe, Gertrude Elizabeth Margaret (1971). *Causality and determination: An inaugural lecture*. CUP Archive.
- Armstrong, David Malet (1986). In defence of structural universals. *Australasian Journal of Philosophy* 64.1, pp. 85–88. DOI: [10.1080/00048408612342261](https://doi.org/10.1080/00048408612342261).

- Armstrong, David Malet (1988). Are quantities relations? A reply to Bigelow and Pargetter. *Philosophical Studies: an International Journal for Philosophy in the Analytic Tradition* 54.3, pp. 305–316.
- (1989). *Universals: An opinionated introduction*. Routledge. DOI: [10.4324/9780429492617](https://doi.org/10.4324/9780429492617).
- (1997). *A world of states of affairs*. Cambridge University Press.
- (2004). *Truth and truthmakers*. Cambridge University Press.
- Arregui, Ana (2005). On the accessibility of possible worlds: The role of tense and aspect. PhD thesis. University of Massachusetts Amherst. URL: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.7358&rep=rep1&type=pdf>.
- (2007). When aspect matters: the case of would-conditionals. *Natural Language Semantics* 15.3, pp. 221–264. DOI: [s11050-007-9019-6](https://doi.org/10.1017/S1105000790196).
- Barreira, Luis and Claudia Valls (2012). *Dynamical systems: An introduction*. Springer. DOI: [10.1007/978-1-4471-4835-7](https://doi.org/10.1007/978-1-4471-4835-7).
- Barrow, John D (2011). Gödel and physics. *Kurt Gödel and the Foundations of Mathematics: Horizons of Truth*. Ed. by Matthias Baaz et al. Cambridge University Press. Chap. 11, pp. 255–276. URL: <https://arxiv.org/pdf/physics/0612253.pdf>.
- Barwise, Jon and Robin Cooper (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy* 4.2, pp. 159–219. ISSN: 1573-0549. DOI: [10.1007/BF00350139](https://doi.org/10.1007/BF00350139).
- Bassi, Itai and Moshe E Bar-Lev (2018). A unified existential semantics for bare conditionals. *Proceedings of Sinn und Bedeutung*. Vol. 21, pp. 125–142. URL: <https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/128>.
- Beaver, David and Brady Clark (2008). *Sense and sensitivity: How focus determines meaning*. Vol. 12. Wiley. DOI: [10.1002/9781444304176](https://doi.org/10.1002/9781444304176).
- Beaver, David I (1995). Presupposition and assertion in dynamic semantics. PhD thesis. The University of Edinburgh. URL: <http://hdl.handle.net/1842/10767>.
- Beckers, Sander (2016). Actual Causation: Definitions and Principles. PhD thesis. KU Leuven. URL: https://limo.libis.be/primo-explore/fulldisplay?docid=LIRIAS1656621&context=L&vid=Lirias&search_scope=Lirias&tab=default_tab&lang=en_US.
- (2021a). Causal sufficiency and actual causation. *Journal of Philosophical Logic* 50.6, pp. 1341–1374. DOI: [10.1007/s10992-021-09632-6](https://doi.org/10.1007/s10992-021-09632-6).
- (2021b). Equivalent Causal Models. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35, pp. 6202–6209. DOI: [10.1609/aaai.v35i7.16771](https://doi.org/10.1609/aaai.v35i7.16771).
- (2021c). The Causal and the Epistemic Conditions for Moral Responsibility. *PhilArchive*.

- Beckers, Sander, Frederick Eberhardt, and Joseph Y Halpern (2020). Approximate causal abstractions. *Uncertainty in Artificial Intelligence*. PMLR, pp. 606–615.
- Beckers, Sander and Joseph Y Halpern (2019). Abstracting causal models. *Proceedings of the aaai conference on artificial intelligence*. Vol. 33, pp. 2678–2685. DOI: [10.1609/aaai.v33i01.33012678](https://doi.org/10.1609/aaai.v33i01.33012678).
- Beckers, Sander and Joost Vennekens (2018). A principled approach to defining actual causation. *Synthese* 195.2, pp. 835–862. DOI: [10.1007/s11229-016-1247-1](https://doi.org/10.1007/s11229-016-1247-1).
- Belnap, Nuel, Michael Perloff, and Ming Xu (2001). *Facing the future: agents and choices in our indeterminist world*. Oxford University Press.
- Benacerraf, Paul (1962). Tasks, super-tasks, and the modern eleatics. *The Journal of Philosophy* 59.24, pp. 765–784. DOI: [10.2307/2023500](https://doi.org/10.2307/2023500).
- Bennett, Jonathan (2003). *A philosophical guide to conditionals*. Oxford University Press.
- Berman, Mitchell N and Guha Krishnamurthi (2021). Bostock was Bogus: Textualism, Pluralism, and Title VII. *Pluralism, and Title VII (February 1, 2021)*. URL: https://scholarship.law.upenn.edu/faculty_scholarship/2577/.
- Berto, Francesco (2018). Aboutness in imagination. *Philosophical Studies* 175.8, pp. 1871–1886. DOI: [10.1007/s11098-017-0937-y](https://doi.org/10.1007/s11098-017-0937-y).
- Biezma, María and Kyle Rawlins (2012). Responding to alternative and polar questions. *Linguistics and Philosophy* 35.5, pp. 361–406. DOI: [10.1007/s10988-012-9123-z](https://doi.org/10.1007/s10988-012-9123-z).
- Binnick, Robert I (1991). *Time and the verb: A guide to tense and aspect*. Oxford University Press.
- Blanchard, Thomas and Jonathan Schaffer (2017). Cause without default. *Making a difference*. Ed. by Huw Price Helen Beebe Christopher Hitchcock. Oxford University Press Oxford, pp. 175–214. DOI: [10.1093/oso/9780198746911.003.0010](https://doi.org/10.1093/oso/9780198746911.003.0010).
- Bontly, Thomas D (2005). Proportionality, causation, and exclusion. *Philosophia* 32.1-4, pp. 331–348.
- Bott, Lewis and Steven Friszon (2022). Salient alternatives facilitate implicatures. *Plos one* 17.3, e0265781.
- Bowerman, Melissa (1986). First steps in acquiring conditionals. *On conditionals*. Cambridge University Press, pp. 285–308.
- Boylan, David and Ginger Schultheis (2017). Strengthening principles and counterfactual semantics. *Proceedings of the 21st Amsterdam Colloquium*. Ed. by Floris Roelofsen Alexandre Cremers Thom van Gessel, pp. 155–164. URL: <https://semanticsarchive.net/Archive/jZiM2FhZ/AC2017-Proceedings.pdf>.
- (2021). How strong is a counterfactual? *The Journal of Philosophy* 118.7, pp. 373–404. DOI: [10.5840/jphil2021118728](https://doi.org/10.5840/jphil2021118728).

- Braham, Matthew and Martin van Hees (2012). An Anatomy of Moral Responsibility. English. *Mind* 121.483, pp. 601–634. DOI: [10.1093/mind/fzs081](https://doi.org/10.1093/mind/fzs081).
- Brand, Myles (1979). Causality. *Current Research in Philosophy of Science*. Ed. by Jr. Peter D. Asquith & Henry E. Kyburg. Philosophy of Science Association, pp. 251–281.
- Breitkopf, Alfred (1978). Axiomatisierung Einiger Begriffe Aus Nelson Goodmans: The Structure of Appearance. *Erkenntnis* 12, pp. 229–247.
- Briggs, Ray (2012). Interventionist counterfactuals. *Philosophical Studies* 160.1, pp. 139–166. DOI: [10.1007/s11098-012-9908-5](https://doi.org/10.1007/s11098-012-9908-5).
- Brody, B. A. (1980). Toward an Aristotelian Theory of Explanation. *Introductory Readings in the Philosophy of Science*. Prometheus Books, pp. 112–123.
- Brouwer, Luitzen EJ (1948). Essentieel-negatieve eigenschappen. *KNAW Proceedings* 51, pp. 963–964.
- Byrne, Ruth (2005). *The Rational Imagination: How people create alternatives to reality*. MIT press. URL: <https://mitpress.mit.edu/books/rational-imagination>.
- Cain, Terrence (2021). Cause for Concern or Cause for Celebration?: Did Bostock v. Clayton County Establish a New Mixed Motive Theory for Title VII Cases and Make It Easier for Plaintiffs to Prove Discrimination Claims? *Seattle UL Rev.* 45, p. 463.
- Campbell, Keith (1981). The metaphysic of abstract particulars. *Midwest studies in philosophy* 6, pp. 477–488. DOI: [j.1475-4975.1981.tb00453.x](https://doi.org/j.1475-4975.1981.tb00453.x).
- Canavotto, Ilaria (2020). Where responsibility takes you: Logics of agency, counterfactuals and norms. PhD thesis. University of Amsterdam. DOI: [10.1007/978-3-031-17111-6](https://doi.org/10.1007/978-3-031-17111-6).
- Cardona, Robert et al. (2021). Constructing Turing complete Euler flows in dimension 3. *Proceedings of the National Academy of Sciences* 118.19, e2026818118. DOI: [10.1073/pnas.2026818118](https://doi.org/10.1073/pnas.2026818118).
- Cariani, Fabrizio (2021). *The Modal Future: A Theory of Future-Directed Thought and Talk*. Cambridge University Press.
- Cariani, Fabrizio and Paolo Santorio (2018). Will done better: Selection semantics, future credence, and indeterminacy. *Mind* 127.505, pp. 129–165. DOI: [10.1093/mind/fzw004](https://doi.org/10.1093/mind/fzw004).
- Carroll, John W. (2020). Laws of Nature. *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University.
- Cartwright, Nancy (1979). Causal laws and effective strategies. *Noûs* 13.4, pp. 419–437. DOI: [10.2307/2215337](https://doi.org/10.2307/2215337).
- Casati, Roberto and Achille C Varzi (1999). *Parts and places: The structures of spatial representation*. MIT press.
- Chalupka, Krzysztof, Frederick Eberhardt, and Pietro Perona (2016). Multi-Level Cause-Effect Systems. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by Arthur Gretton and Christian

- C. Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, pp. 361–369. URL: <https://proceedings.mlr.press/v51/chalupka16.html>.
- (2017). Causal feature learning: an overview. *Behaviormetrika* 44.1, pp. 137–164. DOI: [/0.1007/s41237-016-0008-2](https://doi.org/10.1007/s41237-016-0008-2).
- Chierchia, Gennaro (2010). Mass nouns, vagueness and semantic variation. *Synthese* 174.1, pp. 99–149. DOI: [10.1007/s11229-009-9686-6](https://doi.org/10.1007/s11229-009-9686-6).
- (2013). *Logic in Grammar*. Oxford University Press. DOI: [10.1093/acprof:oso/9780199697977.001.0001](https://doi.org/10.1093/acprof:oso/9780199697977.001.0001).
- Childers, Zachary Witter (2016). "Cause" and affect: evaluative and emotive parameters of meaning among the periphrastic causative verb in English. PhD thesis. University of Texas at Austin. URL: <http://hdl.handle.net/2152/46919>.
- Ciardelli, Ivano (2016). Lifting conditionals to inquisitive semantics. *Semantics and Linguistic Theory*. Vol. 26, pp. 732–752. DOI: [10.3765/salt.v26i0.3811](https://doi.org/10.3765/salt.v26i0.3811).
- Ciardelli, Ivano, Linmin Zhang, and Lucas Champollion (2018). Two switches in the theory of counterfactuals. *Linguistics and Philosophy* 41.6, pp. 577–621. DOI: [10.1007/s10988-018-9232-4](https://doi.org/10.1007/s10988-018-9232-4).
- Cohen, Susannah (2022). Redefining What It Means to Discriminate Because of Sex: Bostock's Equal Protection Implications. *Colum. L. Rev.* 122, p. 407.
- Collins, John, Ned Hall, and Laurie A Paul (2004). Counterfactuals and Causation: History, problems, and prospects. *Causation and Counterfactuals*. MIT Press.
- Condoravdi, Cleo (2002). Temporal Interpretation of Modals: Modals for the Present and for the Past. *The Construction of Meaning*. CSLI Publications, pp. 59–88. URL: <https://semanticsarchive.net/Archive/2JmZTIw0/temp-modals.ps>.
- Cook, Thomas D. and Donald T. Campbell (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin Boston.
- Cooper, Robin (1983). *Quantification and syntactic theory*. Vol. 21. Studies in Linguistics and Philosophy. Springer. DOI: [10.1007/978-94-015-6932-3](https://doi.org/10.1007/978-94-015-6932-3).
- Copley, Bridget (2001). "Be Going To" as a Case of High Aspect. *Semantics and Linguistic Theory*. Vol. 11, pp. 95–113.
- (2008). The plan's the thing: Deconstructing futurate meanings. *Linguistic inquiry* 39.2, pp. 261–274.
- (2009). The semantics of the future. PhD thesis. Massachusetts Institute of Technology. URL: <http://hdl.handle.net/1721.1/8158>.
- Copley, Bridget and Heidi Harley (2015). A force-theoretic framework for event structure. *Linguistics and Philosophy* 38.2, pp. 103–158. DOI: [10.1007/s10988-015-9168-x](https://doi.org/10.1007/s10988-015-9168-x).
- Corkum, Phil (2022). Is 'cause' ambiguous? *Philosophical Studies*, pp. 1–27. DOI: [10.1007/s11098-022-01809-2](https://doi.org/10.1007/s11098-022-01809-2).

- Cotnoir, Aaron (2018). A note on Priest's mereology. *The Australasian Journal of Logic* 15.4, pp. 642–645.
- Cubitt, Toby, David Perez-Garcia, and Michael M Wolf (2015a). Undecidability of the Spectral Gap. *arXiv preprint*. URL: <https://arxiv.org/abs/1502.04573>.
- (2022). Undecidability of the spectral gap. *Forum of Mathematics, Pi*. Vol. 10. Cambridge University Press. DOI: [10.1017/fmp.2021.15](https://doi.org/10.1017/fmp.2021.15).
- Cubitt, Toby S, David Perez-Garcia, and Michael M Wolf (2015b). Undecidability of the spectral gap. *Nature* 528.7581, pp. 207–211. DOI: [10.1038/nature16059](https://doi.org/10.1038/nature16059).
- da Costa, Newton and Francisco Doria (1991). Undecidability and incompleteness in classical mechanics. *International Journal of Theoretical Physics* 30.8, pp. 1041–1073. DOI: [10.1007/BF00671484](https://doi.org/10.1007/BF00671484).
- Dagum, Paul, Adam Galper, and Eric Horvitz (1992). Dynamic network models for forecasting. *Uncertainty in artificial intelligence*, pp. 41–48. DOI: [10.1016/B978-1-4832-8287-9.50010-4](https://doi.org/10.1016/B978-1-4832-8287-9.50010-4).
- Davidson, Donald (1967a). The logical form of action sentences. *The logic of decision and action*. Ed. by N. Rescher. University of Pittsburgh Press.
- (1967b). Truth and meaning. *Philosophy, language, and artificial intelligence*. Ed. by T.L. Rankin J. Kulas J. H. Fetzer. Springer, pp. 93–111. DOI: [10.1007/978-94-009-2727-8_5](https://doi.org/10.1007/978-94-009-2727-8_5).
- (1969). The individuation of events. *Essays in honor of Carl G. Hempel*. Springer, pp. 216–234. DOI: [10.1007/978-94-017-1466-2_11](https://doi.org/10.1007/978-94-017-1466-2_11).
- Dayal, Veneeta (2004). The universal force of free choice. *Linguistic variation yearbook* 4.1, pp. 5–40.
- Deigan, Michael (2020). A plea for inexact truthmaking. *Linguistics and Philosophy* 43.5, pp. 515–536. DOI: [10.1007/s10988-019-09279-2](https://doi.org/10.1007/s10988-019-09279-2).
- Delgrande, James P (1987). A first-order conditional logic for prototypical properties. *Artificial intelligence* 33.1, pp. 105–130. DOI: [10.1016/0004-3702\(87\)90053-1](https://doi.org/10.1016/0004-3702(87)90053-1).
- Dowe, Phil (2000). *Physical causation*. Cambridge University Press.
- Dowty, David R (1979). *Word Meaning and Montague grammar: The semantics of verbs and times in generative semantics and in Montague's PTQ*. Vol. 7. Springer.
- Egré, Paul and Hans Rott (2021). The Logic of Conditionals. *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2021. Metaphysics Research Lab, Stanford University.
- Ehring, Douglas (1987). Causal relata. *Synthese* 73, pp. 319–328. DOI: [10.1007/BF00484745](https://doi.org/10.1007/BF00484745).
- (2009). Causal relata. *The Oxford Handbook of Causation*. Ed. by Helen Beebe, Peter Menzies, and Christopher Hitchcock. Oxford University Press. DOI: [10.1093/oxfordhb/9780199279739.003.0020](https://doi.org/10.1093/oxfordhb/9780199279739.003.0020).

- Einstein, Albert (1916). Die Grundlage der allgemeinen relativitätstheorie. *Annalen der Physik* 49 (7), pp. 769–822. DOI: [10.1002/andp.19163540702](https://doi.org/10.1002/andp.19163540702). URL: <https://einsteinpapers.press.princeton.edu/vol6-trans/158>.
- Ellis, Brian, Frank Jackson, and Robert Pargetter (1977). An objection to possible-world semantics for counterfactual logics. *Journal of Philosophical Logic* 6.1, pp. 355–357. DOI: [10.1007/BF00262069](https://doi.org/10.1007/BF00262069).
- Embry, Brian (2014). Counterfactuals without possible worlds? A Difficulty for Fine’s exact semantics for counterfactuals. *The Journal of Philosophy* 111.5, pp. 276–287. DOI: [10.5840/jphil2014111522](https://doi.org/10.5840/jphil2014111522).
- Enç, Murvet (1996). Tense and modality. *The Handbook of Contemporary Semantic Theory*. Ed. by Shalom Lappin. Blackwell, pp. 345–358.
- Fălăuș, Anamaria and Brenda Laca (2020). Modal-temporal interactions. *The Wiley Blackwell Companion to Semantics*. Ed. by Lisa Matthewson et al. Blackwell. URL: <https://hal.archives-ouvertes.fr/hal-01372977>.
- Fine, Kit (1975a). Critical Notice (review of Lewis 1973b). *Mind* 84.1, pp. 451–458. DOI: [10.1093/mind/LXXXIV.1.451](https://doi.org/10.1093/mind/LXXXIV.1.451).
- (1975b). Vagueness, truth and logic. *Synthese* 30, pp. 265–300. DOI: [10.1007/BF00485047](https://doi.org/10.1007/BF00485047).
- (2002). The varieties of necessity. *Conceivability and possibility*. Ed. by Tamar Szabó Gendler and John Hawthorne. Oxford, pp. 253–281.
- (2012a). A difficulty for the possible worlds analysis of counterfactuals. *Synthese* 189.1, pp. 29–57. DOI: [10.5840/jphil201210938](https://doi.org/10.5840/jphil201210938).
- (2012b). Counterfactuals without possible worlds. *Journal of Philosophy* 109.3, pp. 221–246. DOI: [10.5840/jphil201210938](https://doi.org/10.5840/jphil201210938).
- (2016). Angelic Content. *Journal of Philosophical Logic* 45.2, pp. 199–226. DOI: [10.1007/s10992-015-9371-9](https://doi.org/10.1007/s10992-015-9371-9).
- (2017a). A theory of truthmaker content I: Conjunction, disjunction and negation. *Journal of Philosophical Logic* 46.6, pp. 625–674. DOI: [10.1007/s10992-016-9413-y](https://doi.org/10.1007/s10992-016-9413-y).
- (2017b). Truthmaker Semantics. *A Companion to the Philosophy of Language*. Wiley Blackwell. Chap. 22, pp. 556–577. DOI: [10.1002/9781118972090.ch22](https://doi.org/10.1002/9781118972090.ch22).
- (2021). Constructing the impossible. *Conditionals, Paradox, and Probability: Themes from the Philosophy of Dorothy Edgington*. Ed. by Lee Walters and John Hawthorne. Oxford University Press, pp. 141–163. DOI: [10.1093/oso/9780198712732.003.0009](https://doi.org/10.1093/oso/9780198712732.003.0009).
- von Fintel, Kai (1994). Restrictions on quantifier domains. PhD thesis. University of Massachusetts at Amherst. URL: <https://semanticsarchive.net/Archive/jA3N2IwN/fintel-1994-thesis.pdf>.
- (1997). Bare plurals, bare conditionals, and only. *Journal of Semantics* 14.1, pp. 1–56. DOI: [10.1093/jos/14.1.1](https://doi.org/10.1093/jos/14.1.1).
- (2001a). Conditional strengthening: A case study in implicature. *Unpublished manuscript*. URL: <http://mit.edu/fintel/fintel-2001-condstrengthen.pdf>.

- von Fintel, Kai (2001b). Counterfactuals in a dynamic context. *Ken Hale: A life in language*. Ed. by Michael Kenstowicz. Vol. 36. MIT Press, pp. 123–152. URL: <http://web.mit.edu/fintel/fintel-2001-counterfactuals.pdf>.
- Fodor, Janet D, Jerry A Fodor, and Merrill F Garrett (1975). The Psychological Unreality of Semantic Representations. *Linguistic Inquiry*, pp. 515–531. DOI: [10.4159/harvard.9780674594722.c18](https://doi.org/10.4159/harvard.9780674594722.c18).
- Forrest, Peter (2013). Exemplification and parthood. *Axiomathes* 23.2, pp. 323–341. DOI: [10.1007/s10516-013-9215-6](https://doi.org/10.1007/s10516-013-9215-6).
- Fox, Danny (2003). Implicature calculation, *only*, and lumping: Another look at the puzzle of disjunction. *Yale University Handout*. URL: <https://ling.yale.edu/sites/default/files/files/g2.pdf>.
- (2007). Free choice and the theory of scalar implicatures. *Presupposition and implicature in compositional semantics*. Ed. by Uli Sauerland and Penka Stateva. Palgrave, pp. 71–120. DOI: [10.1057/9780230210752_4](https://doi.org/10.1057/9780230210752_4).
- Fox, Danny and Roni Katzir (2011). On the characterization of alternatives. *Natural language semantics* 19.1, pp. 87–107. DOI: [10.1007/s11050-010-9065-3](https://doi.org/10.1007/s11050-010-9065-3).
- Fox, Danny and Benjamin Spector (2018). Economy and embedded exhaustification. *Natural Language Semantics* 26.1, pp. 1–50. DOI: [10.1007/s11050-017-9139-6](https://doi.org/10.1007/s11050-017-9139-6).
- Freedman, Michael H (1998). P/NP, and the quantum field computer. *Proceedings of the National Academy of Sciences* 95.1, pp. 98–101. DOI: [10.1073/pnas.95.1.98](https://doi.org/10.1073/pnas.95.1.98).
- Fuhrmann, André (1996). *An essay on contraction*. CSLI Publications.
- (1999). When hyperpropositions meet... *Journal of Philosophical Logic* 28, pp. 559–574. DOI: [10.1023/A:1004792327149](https://doi.org/10.1023/A:1004792327149).
- Fusco, Melissa (2019). Sluicing on free choice. *Semantics and Pragmatics* 12, p. 20. DOI: [10.3765/sp.12.20](https://doi.org/10.3765/sp.12.20).
- Gabbay, Dov M (1972). A general theory of the conditional in terms of a ternary operator. *Theoria* 38.3, pp. 97–104. DOI: [10.1111/j.1755-2567.1972.tb00927.x](https://doi.org/10.1111/j.1755-2567.1972.tb00927.x).
- Galles, David and Judea Pearl (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science* 3.1, pp. 151–182. DOI: [10.1023/A:1009602825894](https://doi.org/10.1023/A:1009602825894).
- Gärdenfors, Peter (2000). *Conceptual spaces: The geometry of thought*. MIT press.
- (2014). *The geometry of meaning: Semantics based on conceptual spaces*. MIT press.
- Gasking, Douglas (1955). Causation and recipes. *Mind* 64.256, pp. 479–487. DOI: [10.1093/mind/LXIV.256.479](https://doi.org/10.1093/mind/LXIV.256.479).
- Geiger, Atticus et al. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems* 34, pp. 9574–9586.
- Geis, Michael L. and Arnold M. Zwicky (1971). On invited inferences. *Linguistic inquiry* 2.4, pp. 561–566. URL: www.jstor.org/stable/4177664.

- Geis, Michael Lorenz (1970). Adverbial subordinate clauses in English. PhD thesis. Massachusetts Institute of Technology. URL: <http://hdl.handle.net/1721.1/12971>.
- Gerstenberg, Tobias et al. (2021). A counterfactual simulation model of causal judgment. *Psychological Review*, pp. 936–975. DOI: [10.1037/rev0000281](https://doi.org/10.1037/rev0000281).
- Goldberg, Suzanne B (2011). Discrimination by comparison. *The Yale Law Journal*, pp. 728–812.
- Goldstein, S. et al. (2011). Bell’s theorem. *Scholarpedia* 6.10. revision #91049, p. 8378. DOI: [10.4249/scholarpedia.8378](https://doi.org/10.4249/scholarpedia.8378).
- Goodman, Nelson (1951). *The structure of appearance*. Springer. DOI: [10.1007/978-94-010-1184-6](https://doi.org/10.1007/978-94-010-1184-6).
- Greenland, Sander and Judea Pearl (2011). Causal Diagrams. *International Encyclopedia of Statistical Science*. Springer, pp. 208–216. DOI: [10.1007/978-3-642-04898-2_162](https://doi.org/10.1007/978-3-642-04898-2_162).
- Grimshaw, Jane Barbara (1993). Semantic structure and semantic content in lexical representation. *Words and Structure*. CSLI Publications, pp. 75–89.
- Groenendijk, J.A.G and M.J.B. Stokhof (1984). Studies on the semantics of questions and the pragmatics of answers. PhD thesis. University of Amsterdam. URL: <http://hdl.handle.net/11245/1.392528>.
- Haavelmo, Trygve (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, pp. 1–12. DOI: [10.2307/1905714](https://doi.org/10.2307/1905714).
- Hall, Ned (2000). Causation and the Price of Transitivity. *Journal of Philosophy* 97.4, pp. 198–222. DOI: [10.2307/2678390](https://doi.org/10.2307/2678390).
- (2004). Two concepts of causation. *Causation and counterfactuals*. Ed. by John Collins, Ned Hall, and Paul Laurie. MIT Press, pp. 225–276.
- (2007). Structural equations and causation. *Philosophical Studies* 132.1, pp. 109–136. DOI: [10.1007/s11098-006-9057-9](https://doi.org/10.1007/s11098-006-9057-9).
- Hall, Ned and Laurie A. Paul (2003). Causation and preemption. *Philosophy of Science Today*, pp. 100–130.
- Halpern, Joseph and Christopher Hitchcock (2010). Actual Causation and the Art of Modeling. *Causality, Probability, and Heuristics: A Tribute to Judea Pearl*. College Publications, pp. 383–406.
- Halpern, Joseph Y (2016). *Actual Causality*. MIT Press.
- Halpern, Joseph Y and Judea Pearl (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British journal for the philosophy of science* 56.4, pp. 843–887. DOI: [10.1093/bjps/axi147](https://doi.org/10.1093/bjps/axi147).
- Hamkins, Joel David and Makoto Kikuchi (2016). Set-theoretic mereology. *Logic and Logical Philosophy* 25.3, pp. 285–308. DOI: [10.12775/LLP.2016.007](https://doi.org/10.12775/LLP.2016.007).
- Hart, H. L. A. and Tony Honoré (1959). *Causation in the Law*. Clarendon Press. DOI: [10.1093/acprof:oso/9780198254744.001.0001](https://doi.org/10.1093/acprof:oso/9780198254744.001.0001).
- Hausman, Daniel Murray (2005). Causal relata: Tokens, types, or variables? *Erkenntnis* 63.1, pp. 33–54. DOI: [10.1007/s10670-005-0562-6](https://doi.org/10.1007/s10670-005-0562-6).

- Hawke, Peter (2018). Theories of aboutness. *Australasian Journal of Philosophy* 96.4, pp. 697–723. DOI: [10.1080/00048402.2017.1388826](https://doi.org/10.1080/00048402.2017.1388826).
- Hawking, Stephen (1988). *A brief history of time: from big bang to black holes*. Bantam Dell.
- Heim, Irene (1990). *Only and Lumping*. UCLA and Universität Konstanz.
- Hiddleston, Eric (2005). A Causal Theory of Counterfactuals. *Noûs* 39.4, pp. 632–657. DOI: [10.1111/j.0029-4624.2005.00542.x](https://doi.org/10.1111/j.0029-4624.2005.00542.x).
- Higginbotham, James (2003). Conditionals and compositionality. *Philosophical perspectives* 17, pp. 181–194. DOI: [10.1111/j.1520-8583.2003.00008.x](https://doi.org/10.1111/j.1520-8583.2003.00008.x).
- Hitchcock, Christopher (2012). Events and times: A case study in means-ends metaphysics. *Philosophical Studies* 160.1, pp. 79–96. DOI: [10.1007/s11098-012-9909-4](https://doi.org/10.1007/s11098-012-9909-4).
- Hoek, Daniel (2018). Conversational exculpation. *Philosophical Review* 127.2, pp. 151–196. DOI: [10.1215/00318108-4326594](https://doi.org/10.1215/00318108-4326594).
- Holland, Paul W (1986). Statistics and causal inference. *Journal of the American statistical Association* 81.396, pp. 945–960.
- Hovda, Paul (2009). What is classical mereology? *Journal of Philosophical Logic* 38.1, pp. 55–82. DOI: [10.1007/s10992-008-9092-4](https://doi.org/10.1007/s10992-008-9092-4).
- Huddleston, Rodney and Geoffrey K. Pullum (2004). The classification of finite subordinate clauses. *An International Master of Syntax and Semantics: Papers Presented to Aimo Seppänen on the Occasion of his 75th Birthday*. Ed. by Mats Möller Gunnar Bergh Jennifer Herriman. Vol. 88. Gothenburg Studies in English. Acta Universitatis Gothoburgensis, pp. 103–116.
- Hudson, James L (1975). Logical subtraction. *Analysis* 35.4, pp. 130–135. DOI: [10.1093/analysis/35.4.130](https://doi.org/10.1093/analysis/35.4.130).
- Humberstone, Lloyd (1981). Logical subtraction: Problems and prospects. *Typescript*.
- (2011). *The Connectives*. MIT Press.
- Hume, David (1748). *Philosophical Essays Concerning Human Understanding*. 1st. London: A. Millar.
- Iatridou, Sabine (1991). Topics in conditionals. PhD thesis. Massachusetts Institute of Technology. URL: <http://hdl.handle.net/1721.1/13521>.
- (1993). On the contribution of conditional then. *Natural language semantics* 2.3, pp. 171–199.
- (2021). Grammar matters. *Conditionals, Paradox, and Probability: Themes from the Philosophy of Dorothy Edgington*. Ed. by Lee Walters and John Hawthorne. Oxford University Press. DOI: [10.1093/oso/9780198712732.003.0008](https://doi.org/10.1093/oso/9780198712732.003.0008).
- van Inwagen, Peter (2011). Relational vs. constituent ontologies. *Philosophical perspectives* 25, pp. 389–405. DOI: [10.1111/j.1520-8583.2011.00221.x](https://doi.org/10.1111/j.1520-8583.2011.00221.x).
- Ippolito, Michela (2003). Presuppositions and implicatures in counterfactuals. *Natural language semantics* 11.2, pp. 145–186. DOI: [10.1023/A:1024411924818](https://doi.org/10.1023/A:1024411924818).

- (2006). Semantic composition and presupposition projection in subjunctive conditionals. *Linguistics and Philosophy* 29.6, pp. 631–672. DOI: doi.org/10.1007/s10988-006-9006-2.
- (2008). On the Meaning of Only. *Journal of Semantics* 25.1, pp. 45–91. DOI: [10.1093/jos/ffm010](https://doi.org/10.1093/jos/ffm010).
- (2013). *Subjunctive Conditionals: A Linguistic Analysis*. MIT Press.
- Jacquette, Dale (1990). Wittgenstein and the color incompatibility problem. *History of philosophy quarterly* 7.3, pp. 353–365.
- Jaeger, Robert A (1973). Action and subtraction. *The Philosophical Review* 82.3, pp. 320–329. DOI: [10.2307/2183898](https://doi.org/10.2307/2183898).
- (1976). Logical subtraction and the analysis of action. *Analysis* 36.3, pp. 141–146. DOI: [10.1093/analysis/36.3.141](https://doi.org/10.1093/analysis/36.3.141).
- Jespersen, Otto (1924). *The philosophy of grammar*. George Allen & Unwin.
- Johnson-Laird, P. N. (1982). Formal Semantics and the Psychology of Meaning. *Processes, Beliefs, and Questions: Essays on Formal Semantics of Natural Language and Natural Language Processing*. Ed. by Stanley Peters and Esa Saarinen. Springer, pp. 1–68. DOI: [10.1007/978-94-015-7668-0_1](https://doi.org/10.1007/978-94-015-7668-0_1).
- Kac, Michael B (1972). Clauses of saying and the interpretation of because. *Language*, pp. 626–632.
- Kamali, Beste and Manfred Krifka (2020). Focus and contrastive topic in questions and answers, with particular reference to Turkish. *Theoretical Linguistics* 46.1-2, pp. 1–71.
- Kaminski, Margot E (2019). The right to explanation, explained. *Berkeley Tech. LJ* 34, p. 189. DOI: [10.15779/Z38TD9N83H](https://doi.org/10.15779/Z38TD9N83H).
- Kamp, Hans and Uwe Reyle (1993). Tense and aspect. *From Discourse to Logic*. Springer, pp. 483–689. DOI: [10.1007/978-94-017-1616-1_6](https://doi.org/10.1007/978-94-017-1616-1_6).
- Kant, Immanuel (2004). *Prolegomena to Any Future Metaphysics*. Revised edition. Translated and edited by Gary Hatfield. Cambridge University Press. DOI: [10.1017/CB09780511808517](https://doi.org/10.1017/CB09780511808517).
- Karttunen, Lauri (1971). Counterfactual conditionals. *Linguistic Inquiry* 2.4, pp. 566–569.
- Katz, Jerrold J. (1982). Common sense in semantics. *Notre Dame Journal of Formal Logic* 23.2, pp. 174–218. DOI: [10.1305/ndjfl/1093883626](https://doi.org/10.1305/ndjfl/1093883626).
- Katzir, Roni (2007). Structurally-defined alternatives. *Linguistics and Philosophy* 30.6, pp. 669–690. ISSN: 1573-0549. DOI: [10.1007/s10988-008-9029-y](https://doi.org/10.1007/s10988-008-9029-y).
- Kaufmann, Stefan (2005). Conditional truth and future reference. *Journal of semantics* 22.3, pp. 231–280. DOI: [10.1093/jos/ffh025](https://doi.org/10.1093/jos/ffh025).
- (2013). Causal premise semantics. *Cognitive science* 37.6, pp. 1136–1170. DOI: [10.1111/cogs.12063](https://doi.org/10.1111/cogs.12063).
- Keenan, Edward L and Jonathan Stavi (1986). A semantic characterization of natural language determiners. *Linguistics and philosophy* 9.3, pp. 253–326. DOI: [10.1007/BF00630273](https://doi.org/10.1007/BF00630273).

- Khoo, Justin (2015). On indicative and subjunctive conditionals. *Philosophers' Imprint* 15.32, pp. 1–40. URL: <http://hdl.handle.net/2027/spo.3521354.0015.032>.
- (2017). Backtracking counterfactuals revisited. *Mind* 126.503, pp. 841–910. DOI: [10.1093/mind/fzw005](https://doi.org/10.1093/mind/fzw005).
- (2021a). Coordinating *If*s. *Journal of Semantics* 38.2, pp. 341–361. DOI: [10.1093/jos/ffab006](https://doi.org/10.1093/jos/ffab006).
- (2021b). Disjunctive antecedent conditionals. *Synthese* 198.8, pp. 7401–7430.
- Kim, Jaegwon (1973). Causes and counterfactuals. *The Journal of Philosophy* 70.17, pp. 570–572. DOI: [10.2307/2025312](https://doi.org/10.2307/2025312).
- (1974). Noncausal connections. *Nôûs*, pp. 41–52. DOI: [10.2307/2214644](https://doi.org/10.2307/2214644).
- Klein, Wolfgang (1994). *Time in language*. Routledge.
- Kline, A. David (1980). Are There Cases of Simultaneous Causation? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1, pp. 292–301. DOI: [10.1086/psaprocbienmeetp.1980.1.192573](https://doi.org/10.1086/psaprocbienmeetp.1980.1.192573).
- Koopmans, Tjalling C (1950). When Is an Equation System Complete for Statistical Purposes? *Statistical Inference in Dynamic Economic Models*. Wiley, pp. 393–409. DOI: [10.1017/CB09781139170116.047](https://doi.org/10.1017/CB09781139170116.047).
- Kratzer, Angelika (1977). What ‘must’ and ‘can’ must and can mean. *Linguistics and philosophy* 1.3, pp. 337–355. DOI: [10.1007/BF00353453](https://doi.org/10.1007/BF00353453).
- (1981a). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10.2, pp. 201–216. DOI: [10.1007/BF00248849](https://doi.org/10.1007/BF00248849).
- (1981b). The notional category of modality. *Words, Worlds, and Contexts: New Approaches in Word Semantics*. Berlin: W. de Gruyter. Ed. by Hans-Jürgen Eikmeyer and Hannes Rieser, pp. 39–74. DOI: [10.1515/9783110842524-004](https://doi.org/10.1515/9783110842524-004).
- (1989). An investigation of the lumps of thought. *Linguistics and philosophy*, pp. 607–653. DOI: [10.1007/BF00627775](https://doi.org/10.1007/BF00627775).
- (1990). How specific is a fact? *University of Massachusetts Amherst Handout*. URL: <https://semanticsarchive.net/Archive/mQwZjBj0/facts.pdf>.
- (1998). More structural analogies between pronouns and tenses. *Semantics and linguistic theory*. Vol. 8, pp. 92–110. DOI: [10.3765/salt.v8i0.2808](https://doi.org/10.3765/salt.v8i0.2808).
- (2002). Facts: Particulars or information units? *Linguistics and philosophy* 25, pp. 655–670. DOI: [10.1023/A:1020807615085](https://doi.org/10.1023/A:1020807615085).
- (2012). *Modals and conditionals*. Oxford University Press. DOI: [10.1093/acprof:oso/9780199234684.001.0001](https://doi.org/10.1093/acprof:oso/9780199234684.001.0001).
- Krifka, Manfred (1993). Focus and presupposition in dynamic interpretation. *Journal of semantics* 10.4, pp. 269–300. DOI: [10.1093/jos/10.4.269](https://doi.org/10.1093/jos/10.4.269).
- (1995). The semantics and pragmatics of polarity items. *Linguistic Analysis* 25, pp. 209–57.
- Kügelgen, Julius von, Abdirisak Mohamed, and Sander Beckers (2022). Backtracking Counterfactuals. *arXiv preprint*. DOI: [10.48550/arXiv.2211.00472](https://doi.org/10.48550/arXiv.2211.00472).

- Lassiter, Daniel (2018). Complex sentential operators refute unrestricted Simplification of Disjunctive Antecedents. *Semantics and Pragmatics* 11. DOI: [10.3765/sp.11.9](https://doi.org/10.3765/sp.11.9).
- Leahy, Brian (2011). Presuppositions and antipresuppositions in conditionals. *Semantics and linguistic theory*. Vol. 21, pp. 257–274. DOI: [10.3765/salt.v21i0.2613](https://doi.org/10.3765/salt.v21i0.2613).
- (2018). Counterfactual antecedent falsity and the epistemic sensitivity of counterfactuals. *Philosophical Studies* 175.1, pp. 45–69. DOI: [10.1007/s11098-017-0855-z](https://doi.org/10.1007/s11098-017-0855-z).
- Leśniewski, Stanisław (1927–1931). O podstawach matematyki [On the Foundations of Mathematics], I–V. *Przegląd Filozoficzny*. Volume 30 (1927), 164–206; 31 (1928), 261–291; 32 (1929), 60–101; 33 (1930), 77–105; 34 (1931), 142–170. Abridged English translation by Vito Sinisi 1983.
- Levin, Beth (2005). Semantic Prominence and Argument Realization V: Structuring Event Structure. *MIT Handout*. URL: <https://web.stanford.edu/~bclevin/lisa05evstr.pdf>.
- Lewis, David (1970a). General Semantics. *Semantics of Natural Language*. Ed. by Donald Davidson and Gilbert Harman. Dordrecht: Springer Netherlands, pp. 169–218. DOI: [10.1007/978-94-010-2557-7_7](https://doi.org/10.1007/978-94-010-2557-7_7).
- (1970b). General Semantics. *Synthese* 22, pp. 18–67. DOI: [10.1007/BF00413598](https://doi.org/10.1007/BF00413598).
- (1973a). Causation. *Journal of Philosophy* 70.17, pp. 556–567. DOI: [10.2307/2025310](https://doi.org/10.2307/2025310).
- (1973b). *Counterfactuals*. Wiley-Blackwell.
- (1979). Counterfactual dependence and time’s arrow. *Noûs*, pp. 455–476. DOI: [10.2307/2215339](https://doi.org/10.2307/2215339).
- (1981). Ordering semantics and premise semantics for counterfactuals. *Journal of philosophical logic*, pp. 217–234. DOI: [10.1007/BF00248850](https://doi.org/10.1007/BF00248850).
- (1986). Postscripts to ‘Causation’. *Philosophical Papers Vol. II*. Ed. by David Lewis. Oxford University Press.
- (1988). Relevant implication. *Theoria* 54.3, pp. 161–174. DOI: [10.1111/j.1755-2567.1988.tb00716.x](https://doi.org/10.1111/j.1755-2567.1988.tb00716.x).
- (2000). Causation as Influence. *Journal of Philosophy* 97.4, pp. 182–197. DOI: [10.2307/2678389](https://doi.org/10.2307/2678389).
- Loss, Roberto (2021). Two notions of fusion and the landscape of extensionality. *Philosophical Studies* 178.10, pp. 3443–3463. DOI: [10.1007/s11098-021-01608-1](https://doi.org/10.1007/s11098-021-01608-1).
- Loux, Michael J. (2005). Aristotle on matter, form, and ontological strategy. *Ancient Philosophy* 25.1, pp. 81–123. DOI: [10.5840/ancientphil20052515](https://doi.org/10.5840/ancientphil20052515).
- (2012). What is Constituent Ontology? *Metaphysics: Aristotelian, Scholastic, Analytic*. Ed. by Lukáš Novák Daniel D. Novotný Prokop Sousedík David . De Gruyter. DOI: [10.1515/9783110322446.43](https://doi.org/10.1515/9783110322446.43).

- Lowe, E. J. (2005). Kinds, Essence, and Natural Necessity. *The Four-Category Ontology: A Metaphysical Foundation for Natural Science*. Oxford University Press. DOI: [10.1093/0199254397.003.0009](https://doi.org/10.1093/0199254397.003.0009).
- Machamer, Peter, Lindley Darden, and Carl F Craver (2000). Thinking about mechanisms. *Philosophy of science* 67.1, pp. 1–25.
- Mackie, John L (1974). *The cement of the universe: A study of causation*. Clarendon Press. DOI: [10.1093/0198246420.001.0001](https://doi.org/10.1093/0198246420.001.0001).
- Magri, Giorgio (2009). A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural language semantics* 17.3, pp. 245–297. DOI: [10.1007/s11050-009-9042-x](https://doi.org/10.1007/s11050-009-9042-x).
- Mandelkern, Matthew (2018). Talking about worlds. *Philosophical Perspectives* 32.1, pp. 298–325. DOI: [10.1111/phpe.12112](https://doi.org/10.1111/phpe.12112).
- Maudlin, Tim (2014). *New foundations for physical geometry: the theory of linear structures*. Oxford University Press.
- Mayrhofer, Ralf et al. (2008). Structured correlation from the causal background. *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 30, pp. 303–308. URL: <https://escholarship.org/uc/item/Ofg901v8>.
- McCawley, James David (1973). Syntactic and Logical Arguments for Semantic Structures. *Dimensions of Linguistic Theory*. Ed. by O. Fujimur. TEC Company, Tokyo, pp. 259–376.
- McDonald, Jenn (2022). Apt Causal Models and The Relativity of Actual Causation. *PhilSci Archive*. URL: <http://philsci-archive.pitt.edu/21407/>.
- McDonnell, Neil (2017). Causal exclusion and the limits of proportionality. *Philosophical Studies* 174.6, pp. 1459–1474.
- McHugh, Dean (2018). Diagrammatic Definitions of Causal Claims. *Diagrammatic Representation and Inference*. Ed. by Chapman et al. Springer, pp. 346–354. DOI: [10.1007/978-3-319-91376-6_32](https://doi.org/10.1007/978-3-319-91376-6_32).
- (2020). Are Causes ever too Strong? Downward monotonicity in the causal domain. *Logic, Language, and Meaning: Monotonicity in Logic and Language*. Ed. by D. Deng et al., pp. 125–146. DOI: [10.1007/978-3-662-62843-0](https://doi.org/10.1007/978-3-662-62843-0).
- (2021). Exhaustification in the semantics of *cause* and *because*. *University of Amsterdam Handout*. URL: https://projects.illc.uva.nl/cil/uploaded_files/inlineitem/Exh_in_cause_and_because_MLC_Handout_5_November_202.pdf.
- (2022). Aboutness and Modality. *Proceedings of the 23rd Amsterdam Colloquium*, pp. 194–206. DOI: [10.21942/uva.21739718](https://doi.org/10.21942/uva.21739718).
- (2023). Exhaustification in the semantics of *cause* and *because*. *Glossa GLOWing papers: special collection from Generative Linguistics in the Old World*. DOI: [10.16995/glossa.7663](https://doi.org/10.16995/glossa.7663).
- McKay, Thomas and Peter van Inwagen (1977). Counterfactuals with disjunctive antecedents. *Philosophical Studies* 31.5, pp. 353–356. DOI: [10.1007/BF01873862](https://doi.org/10.1007/BF01873862).

- Mead, C Alden (1964). Possible connection between gravitation and fundamental length. *Physical Review* 135.3B, B849. DOI: [10.1103/PhysRev.135.B849](https://doi.org/10.1103/PhysRev.135.B849).
- Melamed, Yitzhak Y. and Martin Lin (2021). Principle of Sufficient Reason. *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.
- Menzies, Peter (2017). The Problem of Counterfactual Isomorphs. *Making a Difference: Essays on the Philosophy of Causation*. Ed. by Huw Price Helen Beebe Christopher Hitchcock. Oxford University Press, pp. 153–174. DOI: [10.1093/oso/9780198746911.003.0009](https://doi.org/10.1093/oso/9780198746911.003.0009).
- Menzies, Peter and Helen Beebe (2020). Counterfactual Theories of Causation. *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University.
- Menzies, Peter and Huw Price (1993). Causation as a secondary quality. *The British Journal for the Philosophy of Science* 44.2, pp. 187–203. DOI: [10.1093/bjps/44.2.187](https://doi.org/10.1093/bjps/44.2.187).
- Meyer, Marie-Christine (2013). Ignorance and grammar. PhD thesis. MIT.
- Mill, John Stuart (1843). *A system of logic*. Parker.
- Moens, Marc and Mark Steedman (1988). Temporal ontology and temporal reference. *Computational Linguistics* 14, pp. 15–28.
- Moltmann, Friederike (2007). Events, tropes, and truthmaking. *Philosophical Studies* 134.3, pp. 363–403. DOI: [10.1007/s11098-005-0898-4](https://doi.org/10.1007/s11098-005-0898-4).
- Moore, Cristopher (1990). Unpredictability and undecidability in dynamical systems. *Physical Review Letters* 64.20, p. 2354. DOI: [10.1103/PhysRevLett.64.2354](https://doi.org/10.1103/PhysRevLett.64.2354).
- Moore, Michael S. (2005). Causal Relata. *Jahrbuch für Recht und Ethik / Annual Review of Law and Ethics* 13, pp. 589–641. URL: <http://www.jstor.org/stable/43593720>.
- Morgan, J.L. (1969). On Arguing About Semantics. *Papers in Linguistics* 1, pp. 49–70. DOI: [10.1080/08351816909389106](https://doi.org/10.1080/08351816909389106).
- Morreall, John (1979). The evidential use of because. *Research on Language & Social Interaction* 12.1-2, pp. 231–238.
- Moss, Sarah (2013). Subjunctive credences and semantic humility. *Philosophy and Phenomenological Research* 87.2, pp. 251–278. DOI: [10.1111/j.1933-1592.2011.00550.x](https://doi.org/10.1111/j.1933-1592.2011.00550.x).
- Nadathur, Prerna and Sven Lauer (2020). Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa* 5.1, pp. 1–37. DOI: [10.5334/gjgl.497](https://doi.org/10.5334/gjgl.497).
- Nute, Donald (1975). Counterfactuals and the Similarity of Words. *The Journal of Philosophy* 72.21, pp. 773–778. DOI: [10.2307/2025340](https://doi.org/10.2307/2025340).
- (1980a). Conversational scorekeeping and conditionals. *Journal of Philosophical Logic* 9.2, pp. 153–166. DOI: [10.1007/BF00247746](https://doi.org/10.1007/BF00247746).
- (1980b). *Topics in conditional logic*. Springer.

- Nute, Donald (1984). Conditional logic. *Handbook of philosophical logic*. Springer, pp. 387–439. DOI: [10.1007/978-94-009-6259-0_8](https://doi.org/10.1007/978-94-009-6259-0_8).
- Nyhout, Angela and Patricia A Ganea (2019). Mature counterfactual reasoning in 4-and 5-year-olds. *Cognition* 183, pp. 57–66. DOI: [10.1016/j.cognition.2018.10.027](https://doi.org/10.1016/j.cognition.2018.10.027).
- Olson, Eric T. (2017). Properties as Parts of Ordinary Objects. *Being, Freedom, and Method: Themes from the Philosophy of Peter van Inwagen*. Ed. by John A. Keller. Oxford University Press. DOI: [10.1093/acprof:oso/9780198715702.003.0004](https://doi.org/10.1093/acprof:oso/9780198715702.003.0004).
- Otsuka, Jun and Hayato Saigo (2022). On the Equivalence of Causal Models: A Category-Theoretic Approach. *Proceedings of the First Conference on Causal Learning and Reasoning*. Ed. by Bernhard Schölkopf, Caroline Uhler, and Kun Zhang. Vol. 177. Proceedings of Machine Learning Research. PMLR, pp. 634–646. URL: <https://proceedings.mlr.press/v177/otsuka22a.html>.
- Pagin, Peter and Dag Westerståhl (2010a). Compositionality I: Definitions and Variants. *Philosophy Compass* 5.3, pp. 250–264. DOI: [10.1111/j.1747-9991.2009.00228.x](https://doi.org/10.1111/j.1747-9991.2009.00228.x).
- (2010b). Compositionality II: Arguments and problems. *Philosophy Compass* 5.3, pp. 265–282. DOI: [10.1111/j.1747-9991.2009.00229.x](https://doi.org/10.1111/j.1747-9991.2009.00229.x).
- Parsons, Terence (1990). *Events in the Semantics of English*. MIT Press.
- Partee, Barbara (1979). Semantics – mathematics or psychology? *Semantics from different points of view*. Ed. by A. von Stechow R. Bäuerle U. Egli. Springer, pp. 1–14. DOI: [10.1007/978-3-642-67458-7_1](https://doi.org/10.1007/978-3-642-67458-7_1).
- (1984). Compositionality. *Varieties of Formal Semantics, Proceedings of the Fourth Amsterdam Colloquium*. Dordrecht: Foris, pp. 281–312.
- Paul, L. A. (1998). Keeping Track of the Time: Emending the Counterfactual Analysis of Causation. *Analysis* 58.3, pp. 191–198. DOI: [10.1111/1467-8284.00121](https://doi.org/10.1111/1467-8284.00121).
- (2000). Aspect causation. *The Journal of philosophy* 97.4, pp. 235–256. DOI: [10.2307/2678392](https://doi.org/10.2307/2678392).
- Pearl, Judea (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann. DOI: [10.1016/C2009-0-27609-4](https://doi.org/10.1016/C2009-0-27609-4).
- (1995). Causation, action, and counterfactuals. *Computational Learning and Probabilistic Reasoning*. Ed. by A. Gammerman. Wiley, pp. 235–255.
- (2000). *Causality*. Cambridge University Press.
- (2009). *Causality, second edition*. Cambridge University Press.
- Pearl, Judea, Madelyn Glymour, and Nicholas P Jewell (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
- Pearl, Judea and Azaria Paz (1985). Graphoids: Graph-Based Logic for Reasoning about Relevance Relations. *Technical Report 850038 (R-53-L)*, Cognitive Systems Laboratory. University of California, Los Angeles.

- Pears, David (1981). The logical independence of elementary propositions. *Perspectives on the Philosophy of Wittgenstein*. Ed. by Irving Block. MIT Press, pp. 74–84.
- Peirce, C. S. (1867). On an Improvement in Boole's Calculus of Logic. *The Collected Papers of Charles Sanders Peirce*. Ed. by C. Hartshorne and P. Weiss. Vol. III. Harvard University Press, pp. 3–15.
- Piñón, Christopher (1997). Achievements in an event semantics. *Semantics and Linguistic Theory*. Vol. 7, pp. 276–293. DOI: [10.3765/salt.v7i0.2781](https://doi.org/10.3765/salt.v7i0.2781).
- Pinton, Lorenzo (2021). You may like or dislike this thesis, and I do care which: An inquiry into sluicing and free choice. *MSc Thesis, University of Amsterdam*. URL: <https://eprints.illc.uva.nl/id/eprint/1801/1/MoL-2021-11.text.pdf>.
- Poincaré, Henri (1890a). Sur le problème des trois corps et les équations de la dynamique. *Acta mathematica* 13.1, pp. 1–270. DOI: [10.1007/BF02392510](https://doi.org/10.1007/BF02392510).
- (1890b). Sur les équations de la dynamique et le probleme des trois corps. *Acta Mathematica* 13.1–2, p. 270. DOI: [10.1007/BF02392505](https://doi.org/10.1007/BF02392505).
- (1899). *Les méthodes nouvelles de la mécanique céleste*. Vol. 3. Gauthier-Villars et fils.
- Pollock, John L (1976). *Subjunctive reasoning*. Springer. DOI: [10.1007/978-94-010-1500-4](https://doi.org/10.1007/978-94-010-1500-4).
- Powell, Mava Jo (1973). Semantic analysis of Because. PhD thesis. University of British Columbia. DOI: [10.14288/1.0101506](https://doi.org/10.14288/1.0101506).
- Price, Huw (1991). Agency and probabilistic causality. *The British Journal for the Philosophy of Science* 42.2, pp. 157–176. DOI: [10.1093/bjps/42.2.157](https://doi.org/10.1093/bjps/42.2.157).
- (1992). Agency and causal asymmetry. *Mind* 101.403, pp. 501–520. DOI: [10.1093/mind/101.403.501](https://doi.org/10.1093/mind/101.403.501).
- Priest, Graham (2008). *An introduction to non-classical logic: From if to is*. Cambridge University Press.
- Prior, Arthur Norman (1976). It was to be. *Papers in semantics and ethics*. Ed. by Peter Geach and Anthony Kenny. Duckworth, pp. 97–108.
- Pullum, Geoffrey K. (2009). Lexical categorization in English dictionaries and traditional grammars. *Zeitschrift für Anglistik und Amerikanistik* 57.3, pp. 255–273.
- (2014). Because syntax. *Language Log*. URL: <https://languagelog.ldc.upenn.edu/nll/?p=9494>.
- Pyykkö, Pekka and Jean Paul Desclaux (1979). Relativity and the periodic system of elements. *Accounts of Chemical Research* 12.8, pp. 276–281. DOI: [10.1021/ar50140a002](https://doi.org/10.1021/ar50140a002).
- Quine, W.V.O. (1950). *Methods of Logic*. New York: Holt, Rinehart and Winston.
- Quine, Willard Van Orman (1960). *Word and object*. MIT press.
- Ramsey, Frank P. (1929a). General Propositions and Causality. *Manuscript*. URL: <http://www.dspace.cam.ac.uk/handle/1810/194722>.

- Ramsey, Frank P. (1929b). General Propositions and Causality. *F. P. Ramsey: Philosophical Papers*. Ed. by D. H. Mellor. Cambridge University Press. Chap. 7B, pp. 145–163.
- Rawlins, Kyle (2013). (Un)conditionals. *Natural language semantics* 21.2, pp. 111–178. DOI: [10.1007/s11050-012-9087-0](https://doi.org/10.1007/s11050-012-9087-0).
- Reichenbach, Hans (1947). *Elements of symbolic logic*. Dover Publications.
- Reif, John H., J Doug Tygar, and Akitoshi Yoshida (1994). Computability and complexity of ray tracing. *Discrete & computational geometry* 11.3, pp. 265–288.
- Reilly, Judy Snitzer (1982). *The acquisition of conditionals in English*. University of California, Los Angeles.
- Repp, Sophie and Katharina Spalek (2021). The role of alternatives in language. *Frontiers in Communication* 6, p. 682009.
- Review, Harvard Law (2017). Rethinking actual causation in tort law. *Harvard Law Review* 130, pp. 2163–2182. URL: <https://harvardlawreview.org/2017/06/rethinking-actual-causation-in-tort-law/>.
- Rice, Hugh (1999). David Lewis’s awkward cases of redundant causation. *Analysis* 59.3, pp. 157–164.
- Richardson, Daniel (1969). Some undecidable problems involving elementary functions of a real variable. *The Journal of Symbolic Logic* 33.4, pp. 514–520. DOI: [10.2307/2271358](https://doi.org/10.2307/2271358).
- Roberts, Craige (1989). Modal subordination and pronominal anaphora in discourse. *Linguistics and philosophy* 12.6, pp. 683–721. DOI: [10.1007/BF00632602](https://doi.org/10.1007/BF00632602).
- Romoli, Jacopo (2012). Soft but strong. Neg-raising, soft triggers, and exhaustification. PhD thesis. Harvard University. URL: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:9909638>.
- (2015). The Presuppositions of Soft Triggers are Obligatory Scalar Implicatures. *Journal of Semantics* 32.2, pp. 173–219. DOI: [10.1093/jos/fft017](https://doi.org/10.1093/jos/fft017).
- van Rooij, Robert (2006). Free choice counterfactual donkeys. *Journal of Semantics* 23.4, pp. 383–402.
- van Rooij, Robert and Katrin Schulz (2004). Exhaustive Interpretation of Complex Sentences. *Journal of Logic, Language and Information* 13.4, pp. 491–519. DOI: [10.1007/s10849-004-2118-6](https://doi.org/10.1007/s10849-004-2118-6).
- (2007). Only: Meaning and Implicatures. *Questions in dynamic semantics*. Brill, pp. 193–223. DOI: [10.1163/9780080470993_010](https://doi.org/10.1163/9780080470993_010).
- Rooth, Mats (1985). Association with focus. PhD thesis. University of Massachusetts, Amherst. URL: [url=https%5C%3A%5C%2F%5C%2Fecommons.cornell.edu%5C%2Fbitstream%5C%2Fhandle%5C%2F1813%5C%2F28568%5C%2FRooth-1985-PhD.pdf&usg=A0vVaw2X_mAUuPyshpTVYY-Q3mK_](https://www.commonscornell.edu/bitstream/handle/1813/28568/2/Rooth-1985-PhD.pdf&usg=A0vVaw2X_mAUuPyshpTVYY-Q3mK_).
- (1999). Association with focus or association with presupposition? *Focus: Linguistic, cognitive, and computational perspectives*. Ed. by Peter Bosch and Rob van der Sandt. Cambridge University Press, pp. 232–246.

- Ross, Hayley (2020). The Falsity of the Consequent in Contrastive Conditionals. *Brandeis University Masters Thesis*. URL: <https://hdl.handle.net/10192/37528>.
- Rubenstein, Paul K et al. (2017). Causal consistency of structural equation models. *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*. URL: <https://www.auai.org/uai2017/accepted.php>.
- Rullmann, Hotze and Lisa Matthewson (2018). Towards a theory of modal-temporal interaction. *Language* 94.2, pp. 281–331. DOI: [10.1353/lan.2018.0018](https://doi.org/10.1353/lan.2018.0018).
- Salmon, Wesley C (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.
- (1994). Causality without counterfactuals. *Philosophy of Science* 61.2, pp. 297–312. DOI: [10.1086/289801](https://doi.org/10.1086/289801).
- Santorio, Paolo (2014). Filtering semantics for counterfactuals: Bridging causal models and premise semantics. *Semantics and Linguistic Theory*. Vol. 24, pp. 494–513. DOI: [10.3765/salt.v24i0.2430](https://doi.org/10.3765/salt.v24i0.2430).
- (2018). Alternatives and truthmakers in conditional semantics. *The Journal of Philosophy* 115.10, pp. 513–549. DOI: [10.5840/jphil20181151030](https://doi.org/10.5840/jphil20181151030).
- (2019). Interventions in premise semantics. *Philosophers' Imprint*. URL: <http://hdl.handle.net/2027/spo.3521354.0019.001>.
- Sartorio, Carolina (2004). How to be responsible for something without causing it. *Philosophical Perspectives* 18, pp. 315–336.
- (2005). Causes As Difference-Makers. *Philosophical Studies* 123.1, pp. 71–96. DOI: [10.1007/s11098-004-5217-y](https://doi.org/10.1007/s11098-004-5217-y).
- (2006). On causing something to happen in a certain way without causing it to happen. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 129.1, pp. 119–136. DOI: [0.1007/s11098-005-3023-9](https://doi.org/10.1007/s11098-005-3023-9).
- Schaffer, Jonathan (2000). Trumping Preemption. *Journal of Philosophy* 97.4, pp. 165–181. DOI: [10.2307/2678388](https://doi.org/10.2307/2678388).
- (2005). Contrastive causation. *The Philosophical Review* 114.3, pp. 327–358. DOI: [10.1215/00318108-114-3-327](https://doi.org/10.1215/00318108-114-3-327).
- (2010). Contrastive causation in the law. *Legal Theory* 16.4, pp. 259–297. DOI: [10.1017/S1352325210000224](https://doi.org/10.1017/S1352325210000224).
- (2016). The Metaphysics of Causation. *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2016. Metaphysics Research Lab, Stanford University.
- Schulz, Katrin (2007). Minimal models in semantics and pragmatics: Free choice, exhaustivity, and conditionals. PhD thesis. Institute for Logic, Language and Computation, University of Amsterdam. URL: <https://www.illc.uva.nl/Research/Publications/Dissertations/DS-2007-04.text.pdf>.
- (2011). “If you’d wiggled A, then B would’ve changed”. *Synthese* 179.2, pp. 239–251. DOI: [10.1007/s11229-010-9780-9](https://doi.org/10.1007/s11229-010-9780-9).

- Schulz, Katrin and Robert van Rooij (2006). Pragmatic Meaning and Non-monotonic Reasoning: The Case of Exhaustive Interpretation. *Linguistics and Philosophy* 29.2, pp. 205–250. DOI: [10.1007/s10988-005-3760-4](https://doi.org/10.1007/s10988-005-3760-4).
- Seelau, Eric P et al. (1995). Counterfactual constraints. *What Might Have Been: The Social Psychology of Counterfactual Thinking*. Erlbaum.
- Shapiro, Larry and Elliott Sober (2012). Against proportionality. *Analysis* 72.1, pp. 89–93.
- Siegelmann, Hava T, Bill G Horne, and C Lee Giles (1997). Computational capabilities of recurrent NARX neural networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 27.2, pp. 208–215. DOI: [10.1109/3477.558801](https://doi.org/10.1109/3477.558801).
- Simons, Peter (1987). *Parts: A study in ontology*. Oxford University Press.
- (2020). Stanisław Leśniewski. *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2020. Metaphysics Research Lab, Stanford University.
- Sinisi, Vito (1983). Leśniewski's foundations of mathematics. *Topoi* 3.1, pp. 3–52. DOI: [10.1007/BF00139700](https://doi.org/10.1007/BF00139700).
- Skyrms, Brian (1980). *Causal Necessity: a pragmatic investigation of the necessity of laws*. Yale University Press.
- Slote, Michael A (1978). Time in counterfactuals. *The Philosophical Review* 87.1, pp. 3–27. DOI: [10.2307/2184345](https://doi.org/10.2307/2184345).
- Smolin, Lee (2013). *Time reborn*. Houghton Mifflin Harcourt.
- Spector, Benjamin (2003). Scalar implicatures: Exhaustivity and Gricean reasoning. *Proceedings of the Eighth ESSLI Student Session*. Ed. by Balder ten Cate, pp. 277–288.
- (2007). Scalar implicatures: Exhaustivity and gricean reasoning. *Questions in dynamic semantics*. Brill, pp. 225–249. DOI: doi.org/10.1163/9780080470993_011.
- (2016). Comparing exhaustivity operators. *Semantics and Pragmatics* 9, pp. 1–33. DOI: [10.3765/sp.9.11](https://doi.org/10.3765/sp.9.11).
- Spirtes, Peter, Clark Glymour, and Richard Scheines (1993). *Causation, prediction, and search*. MIT Press. DOI: [10.1007/978-1-4612-2748-9](https://doi.org/10.1007/978-1-4612-2748-9).
- Spohn, Wolfgang (2001). Bayesian Nets Are All There Is to Causal Dependence. *Stochastic Causality*. Ed. by Domenico Costantini, Maria Carla Galavotti, and Patrick Suppes. CSLI Publications. Chap. 9.
- (2009). Bayesian Nets Are All There Is to Causal Dependence. *Causation, Coherence, and Concepts: A Collection of Essays*. Springer. Chap. 4, pp. 99–111. DOI: [10.1007/978-1-4020-5474-7_4](https://doi.org/10.1007/978-1-4020-5474-7_4).
- Sportiche, Dominique, Hilda Koopman, and Edward Stabler (2013). *An introduction to syntactic analysis and theory*. John Wiley & Sons.
- Stalnaker, Robert (1968). A theory of conditionals. *Ifs*. Ed. by William L. Harper, Robert Stalnaker, and Glenn Pearce. Springer, pp. 41–55. DOI: [10.1007/978-94-009-9117-0_2](https://doi.org/10.1007/978-94-009-9117-0_2).
- (1984). *Inquiry*. MIT Press.

- (1994). What is a nonmonotonic consequence relation? *Fundamenta Informaticae* 21.1-2, pp. 7–21. DOI: [10.3233/FI-1994-21121](https://doi.org/10.3233/FI-1994-21121).
- Stanley, Jason and Zoltan Szabó (2000). On quantifier domain restriction. *Mind & Language* 15.2-3, pp. 219–261. DOI: [doi:10.1111/1468-0017.00130](https://doi.org/10.1111/1468-0017.00130).
- Starr, William B. (2014). A Uniform Theory of Conditionals. *Journal of Philosophical Logic* 43.6, pp. 1019–1064. DOI: [10.1007/s10992-013-9300-8](https://doi.org/10.1007/s10992-013-9300-8).
- von Stechow, Arnim (1995). Lexical decomposition in syntax. *Lexical Knowledge in the Organization of Language*. John Benjamins, pp. 81–118.
- Suppes, Patrick (1970). *A Theory of Probabilistic Causality*. Amsterdam: North-Holland Publishing Company.
- Talmy, Leonard (1988). Force Dynamics in Language and Cognition. *Cognitive Science* 12.1, pp. 49–100. DOI: [10.1207/s15516709cog1201_2](https://doi.org/10.1207/s15516709cog1201_2).
- Tao, Terence (2016). Finite time blowup for an averaged three-dimensional Navier-Stokes equation. *Journal of the American Mathematical Society* 29.3, pp. 601–674. DOI: [10.1090/jams/838](https://doi.org/10.1090/jams/838).
- (2017). On the universality of potential well dynamics. *Dynamics of Partial Differential Equations* 14.3, pp. 219–238. DOI: [10.4310/DPDE.2017.v14.n3.a1](https://doi.org/10.4310/DPDE.2017.v14.n3.a1).
- Tarski, A, A Mostowski, and RM Robinson (1953). *Undecidable theories*. North Holland.
- Taylor, Richard (1966). *Action and Purpose*. Prentice-Hall.
- Temkin, Jennifer (2002). *Rape and the legal process*. Oxford University Press.
- Thomas, William and Ashwini Deo (2020). The interaction of just with modified scalar predicates. *Proceedings of Sinn und Bedeutung*. Vol. 24, pp. 354–372. DOI: [10.18148/sub/2020.v24i2.902](https://doi.org/10.18148/sub/2020.v24i2.902).
- Thomason, Richmond (1970). Indeterminist time and truth-value gaps. *Theoria* 36.3, pp. 264–281. DOI: [10.1111/j.1755-2567.1970.tb00427.x](https://doi.org/10.1111/j.1755-2567.1970.tb00427.x).
- (2014). Formal semantics for causal constructions. *Causation in Grammatical Structures*. Oxford University Press. DOI: [10.1093/acprof:oso/9780199672073.003.0003](https://doi.org/10.1093/acprof:oso/9780199672073.003.0003).
- Thomason, Richmond and Anil Gupta (1980). A theory of conditionals in the context of branching time. *Ifs*. Springer, pp. 299–322. DOI: [10.1007/978-94-009-9117-0_15](https://doi.org/10.1007/978-94-009-9117-0_15).
- Thomson, James F. (1954). Tasks and Super-Tasks. *Analysis* 15.1, pp. 1–13. DOI: [10.2307/3326643](https://doi.org/10.2307/3326643).
- Tichý, Pavel (1976). A counterexample to the Stalnaker-Lewis analysis of counterfactuals. *Philosophical Studies* 29.4, pp. 271–273. DOI: [10.1007/BF00411887](https://doi.org/10.1007/BF00411887).
- Tulling, Maxime and Ailís Cournane (2019). The role of “fake” past tense in acquiring counterfactuals. *Proceedings of the 2019 Amsterdam Colloquium*. URL: https://archive.illc.uva.nl/AC/AC2019/uploaded_files/inlineitem/Tulling_and_Cournane_The_role_of_fake_past_tense_in.pdf.

- Turing, Alan Mathison (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* 42.230–265, p. 5. DOI: [10.1112/plms/s2-42.1.230](https://doi.org/10.1112/plms/s2-42.1.230).
- Van Canegem-Ardijns, Ingrid (2010). The indefeasibility of the inference that *if not-A, then not-C*. *Journal of Pragmatics* 42.1, pp. 1–15. DOI: [10.1016/j.pragma.2009.05.005](https://doi.org/10.1016/j.pragma.2009.05.005).
- Van Tiel, Bob and Walter Schaeken (2017). Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive science* 41, pp. 1119–1154.
- Veltman, Frank (1976). Prejudices, presuppositions and the theory of counterfactuals. *Proceedings of the 1st Amsterdam Colloquium [=Amsterdam papers in formal grammar]* 1. Ed. by J. Groenendijk M., pp. 248–281. URL: <https://staff.fnwi.uva.nl/f.j.m.m.veltman/papers/Veltman1976.pdf>.
- (2005). Making counterfactual assumptions. *Journal of Semantics* 22.2, pp. 159–180. DOI: [10.1093/jos/ffh022](https://doi.org/10.1093/jos/ffh022).
- Vendler, Zeno (1957). Verbs and times. *The philosophical review* 66.2, pp. 143–160. DOI: [10.2307/2182371](https://doi.org/10.2307/2182371).
- (1967). Facts and Events. *Linguistics in philosophy*. Cornell University Press, pp. 122–146. DOI: [10.7591/9781501743726-006](https://doi.org/10.7591/9781501743726-006).
- Von Wright, Georg Henrik (1971). *Explanation and understanding*. Cornell University Press.
- Vostrikova, Ekaterina (2018). On the similarity between unless and only-if-not. *Proceedings of Sinn und Bedeutung*. Vol. 21, pp. 1271–1288. URL: <https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/198>.
- Vredenburg, Kate (2022). The right to explanation. *Journal of Political Philosophy* 30.2, pp. 209–229. DOI: [10.1111/jopp.12262](https://doi.org/10.1111/jopp.12262).
- Wachter, Sandra, Brent Mittelstadt, and Chris Russell (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology* 31, p. 841. URL: <https://heinonline.org/HOL/P?h=hein.journals/hjlt31&i=859>.
- Warmbröd, Ken (1981). Counterfactuals and substitution of equivalent antecedents. *Journal of Philosophical Logic* 10.2, pp. 267–289. DOI: [10.1007/BF00248853](https://doi.org/10.1007/BF00248853).
- Weslake, Brad (2013). Proportionality, contrast and explanation. *Australasian Journal of Philosophy* 91.4, pp. 785–797.
- (2015). A Partial Theory of Actual Causation. *British Journal for the Philosophy of Science*.
- Willer, Malte (2018). Simplifying with Free Choice. *Topoi* 37.3, pp. 379–392. DOI: [10.1007/s11245-016-9437-5](https://doi.org/10.1007/s11245-016-9437-5).
- Williams, Donald C. (1953). On the elements of being: II. *The review of metaphysics* 7.2, pp. 171–192.
- Wittgenstein, Ludwig (1977). *Remarks on colour*. Edited by G. E. M. Anscombe, translated by Linda Schättle. Blackwell.

- Wolff, Phillip (2007). Representing causation. *Journal of experimental psychology: General* 136.1, p. 82. DOI: [10.1037/0096-3445.136.1.82](https://doi.org/10.1037/0096-3445.136.1.82).
- Wolfram, Stephen (1985). Undecidability and intractability in theoretical physics. *Physical Review Letters* 54.8, p. 735. DOI: [10.1103/PhysRevLett.54.735](https://doi.org/10.1103/PhysRevLett.54.735).
- Woodward, James (2016a). Causation and Manipulability. *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.
- (2016b). The problem of variable choice. *Synthese* 193, pp. 1047–1072. DOI: [10.1007/s11229-015-0810-5](https://doi.org/10.1007/s11229-015-0810-5).
- Wright, Richard (1985). Causation in tort law. *California Law Review* 73.6, pp. 1735–1828. DOI: [10.2307/3480373](https://doi.org/10.2307/3480373).
- (2011). The NESS account of natural causation: a response to criticisms. *Perspectives on Causation*. Ed. by Richard Goldberg. Hart Publishing, pp. 13–66.
- Wright, Sewall (1921). Correlation and causation. *Journal of Agricultural Research* 20.7, pp. 557–580. URL: <https://handle.nal.usda.gov/10113/IND43966364>.
- Xiang, Yimei (2016). Mandarin particle *dou*: A pre-exhaustification exhaustifier. *Empirical issues in syntax and semantics* 11, pp. 275–304. URL: http://www.cssp.cnrs.fr/eiss11/eiss11_xiang.pdf.
- Yablo, Stephen (1992). Mental causation. *The Philosophical Review* 101.2, pp. 245–280.
- (2004). Advertisement for a Sketch of an Outline of a Proto-Theory of Causation. *Causation and Counterfactuals*. Ed. by Ned Hall, L. A. Paul, and John Collins. MIT Press, pp. 119–137.
- (2014). *Aboutness*. Princeton University Press. DOI: [10.1515/9781400845989](https://doi.org/10.1515/9781400845989).
- Yalcin, Seth (2010). Probability operators. *Philosophy Compass* 5.11, pp. 916–937. DOI: [10.1111/j.1747-9991.2010.00360.x](https://doi.org/10.1111/j.1747-9991.2010.00360.x).
- Yang, Eric (2018). Defending constituent ontology. *Philosophical Studies* 175.5, pp. 1207–1216. DOI: [10.1007/s11098-017-0903-8](https://doi.org/10.1007/s11098-017-0903-8).
- Younes, Nadja and Ulf-Dietrich Reips (2019). Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms. *PloS one* 14.3. DOI: [10.1371/journal.pone.0213554](https://doi.org/10.1371/journal.pone.0213554).
- Zhang, Zheng et al. (2023). Scalar Implicature is Sensitive to Contextual Alternatives. *Cognitive Science* 47.2, e13238.
- Zimmermann, Thomas Ede (2000). Free choice disjunction and epistemic possibility. *Natural language semantics* 8.4, pp. 255–290. DOI: [10.1023/A:1011255819284](https://doi.org/10.1023/A:1011255819284).

Samenvatting

Causaliteit en Modaliteit: Modellen en Betekenissen

Dit proefschrift heeft tot doel twee vragen te beantwoorden over causale beweringen (zoals zinnen die *cause* of *because* bevatten). Ten eerste de modelleervraag: wat voor soort informatie gebruiken we wanneer we beoordelen dat een causale bewering geldt? Ten tweede de betekenisvraag: onder welke voorwaarden oordelen we dat een causale bewering waar is?

Ons antwoord op de modelleervraag is dat een causaal model tijd, deel-geheel-structuren en algemeen geldige mogelijkheid moet bevatten. Het model geeft scenario's weer als verlengd in de tijd, waar elk moment in de tijd een mereologische structuur heeft (de mereologische structuur vertelt ons bijvoorbeeld dat de toestand van Amsterdam deel uitmaakt van de toestand van Nederland). Het begrip algemeen geldige mogelijkheid specificeert voor welke werelden het algemeen geldig is dat ze mogelijk en onmogelijk zijn; met andere woorden, welke werelden aan de wetten voldoen en welke niet. Daarnaast moet het model ook twee taalgerelateerde componenten bevatten. Voor elke zin moet het model ons vertellen over welke delen van de wereld het gaat, en in welke werelden de zin waar is.

We laten zien dat dit modelleringskader strikt algemener is dan een populair alternatief, dat van structurele causale modellen. Elk structureel causaal model kan in ons raamwerk worden weergegeven, en daarom geldt dat ons raamwerk altijd een scenario kan weergeven, als structurele causale modellen dat ook kunnen. Het omgekeerde geldt echter niet. Er zijn enkele scenario's die het door ons voorgestelde model kan weergeven, maar die structurele causale modellen niet kunnen weergeven.

We gebruiken deze componenten om te analyseren hoe mensen hypothetische alternatieven voor de werkelijkheid construeren. Algemeen wordt namelijk aangenomen dat de waarheid van een oorzakelijke bewering niet alleen afhangt van wat er in de werkelijke wereld gebeurt, maar ook van wat er in sommige hypothetische scenario's gebeurt. Als we bijvoorbeeld *Alice missed her flight because*

she got stuck in traffic evalueren, stellen we ons scenario's voor waarin ze vastzit in het verkeer en scenario's waarin dit niet zo is, en vergelijken we wat er in elk scenario gebeurt.

Vervolgens analyseren we de semantiek van *cause* en *because* in termen van twee relaties: toereikendheid en productie. De betekenis van *cause* en *because* is een combinatie van deze relaties: *C causes E* en *E because C* zijn waar slechts als *C* waar is, en als *C* voldoende is om *E* te produceren, maar de ontkenning van *C* niet.

Abstract

Causation and Modality: Models and Meanings

This thesis aims to answer two questions about causal claims (such as sentences containing *cause* or *because*). Firstly, the modelling question: what kind of information do we use when we judge that a causal claim holds? Secondly, the meaning question: under what conditions do we judge that a causal claim is true?

Our answer to the modelling question is that a causal model must contain time, part-whole structure, and nomic possibility. The model represents scenarios as extended in time, with each moment in time having a mereological structure (the mereological structure tell us, for example, that the state of Amsterdam is part the state of the Netherlands). The notion of nomic possibility specifies which worlds are nomically possible and which worlds are nomically impossible; in other words, which worlds satisfy the laws and which do not. In addition, the model must also contain two language-related components. For each sentence, the model must tell us what parts of the world it is about, and in which worlds the sentence is true.

We show that this this modelling framework is strictly more general than a popular alternative, that of structural causal models. Every structural causal model can be represented into our framework, and therefore every scenario that structural causal models can represent our framework can represent too. However, the converse does not hold. There are some scenarios that our proposed model can represent which structural causal models cannot.

We use these components to analyse how people construct hypothetical alternatives to reality. For it is commonly thought that the truth of a causal claim depends not only on what goes on in the actual world, but on what happens in some hypothetical scenarios as well. For example, when we evaluate *Alice missed her flight because she got stuck in traffic*, we imagine scenarios where she is stuck in traffic and scenarios where she is not, and compare what happens in each.

We then analyse the semantics of *cause* and *because* in terms of two relations: sufficiency and production. The meaning of *cause* and *because* is a blend of these relations: *C cause E* and *E because C* are true just in case *C* is true, and *C* is sufficient to produce *E* but *C*'s negation is not.

Titles in the ILLC Dissertation Series:

ILLC DS-2018-03: **Corina Koolen**

Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**

Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**

Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**

Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**

Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**

Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**

Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**

Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**

New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**

Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**

Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**

What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**

Quantum logics for expressing and proving the correctness of quantum programs

- ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications
- ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic
- ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles
- ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry
- ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory
- ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding
- ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices
- ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity
- ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization
- ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes
- ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing
- ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning
- ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them
- ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing
- ILLC DS-2020-10: **Arnold Kochari**
Perceiving and communicating magnitudes: Behavioral and electrophysiological studies

- ILLC DS-2020-11: **Marco Del Tredici**
Linguistic Variation in Online Communities: A Computational Perspective
- ILLC DS-2020-12: **Bastiaan van der Weij**
Experienced listeners: Modeling the influence of long-term musical exposure on rhythm perception
- ILLC DS-2020-13: **Thom van Gessel**
Questions in Context
- ILLC DS-2020-14: **Gianluca Grilletti**
Questions & Quantification: A study of first order inquisitive logic
- ILLC DS-2020-15: **Tom Schoonen**
Tales of Similarity and Imagination. A modest epistemology of possibility
- ILLC DS-2020-16: **Ilaria Canavotto**
Where Responsibility Takes You: Logics of Agency, Counterfactuals and Norms
- ILLC DS-2020-17: **Francesca Zaffora Blando**
Patterns and Probabilities: A Study in Algorithmic Randomness and Computable Learning
- ILLC DS-2021-01: **Yfke Dulek**
Delegated and Distributed Quantum Computation
- ILLC DS-2021-02: **Elbert J. Booij**
The Things Before Us: On What it Is to Be an Object
- ILLC DS-2021-03: **Seyyed Hadi Hashemi**
Modeling Users Interacting with Smart Devices
- ILLC DS-2021-04: **Sophie Arnoult**
Adjunction in Hierarchical Phrase-Based Translation
- ILLC DS-2021-05: **Cian Guilfoyle Chartier**
A Pragmatic Defense of Logical Pluralism
- ILLC DS-2021-06: **Zoi Terzopoulou**
Collective Decisions with Incomplete Individual Opinions
- ILLC DS-2021-07: **Anthia Solaki**
Logical Models for Bounded Reasoners
- ILLC DS-2021-08: **Michael Sejr Schlichtkrull**
Incorporating Structure into Neural Models for Language Processing

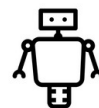
- ILLC DS-2021-09: **Taichi Uemura**
Abstract and Concrete Type Theories
- ILLC DS-2021-10: **Levin Hornischer**
Dynamical Systems via Domains: Toward a Unified Foundation of Symbolic and Non-symbolic Computation
- ILLC DS-2021-11: **Sirin Botan**
Strategyproof Social Choice for Restricted Domains
- ILLC DS-2021-12: **Michael Cohen**
Dynamic Introspection
- ILLC DS-2021-13: **Dazhu Li**
Formal Threads in the Social Fabric: Studies in the Logical Dynamics of Multi-Agent Interaction
- ILLC DS-2022-01: **Anna Bellomo**
Sums, Numbers and Infinity: Collections in Bolzano's Mathematics and Philosophy
- ILLC DS-2022-02: **Jan Czakowski**
Post-Quantum Security of Hash Functions
- ILLC DS-2022-03: **Sonia Ramotowska**
Quantifying quantifier representations: Experimental studies, computational modeling, and individual differences
- ILLC DS-2022-04: **Ruben Brokkelkamp**
How Close Does It Get?: From Near-Optimal Network Algorithms to Suboptimal Equilibrium Outcomes
- ILLC DS-2022-05: **Lwenn Bussière-Caraes**
No means No! Speech Acts in Conflict
- ILLC DS-2023-01: **Subhasree Patro**
Quantum Fine-Grained Complexity
- ILLC DS-2023-02: **Arjan Cornelissen**
Quantum multivariate estimation and span program algorithms
- ILLC DS-2023-03: **Robert Paßmann**
Logical Structure of Constructive Set Theories
- ILLC DS-2023-04: **Samira Abnar**
Inductive Biases for Learning Natural Language

This thesis aims to answer two questions about causal claims (such as sentences containing *cause* or *because*). Firstly, the **modelling question**: what kind of information do we use when we judge that a causal claim holds? Secondly, the **meaning question**: under what conditions do we judge that a causal claim is true?

Our answer to the modelling question is that a causal model must contain time, part-whole structure, and nomic possibility. The model represents scenarios as extended in time, with each moment in time having a mereological structure (the mereological structure tells us, for example, that the state of Amsterdam is part the state of the Netherlands). The notion of nomic possibility specifies which worlds are nomically possible and which worlds are nomically impossible; in other words, which worlds satisfy the laws and which do not. In addition, the model must also contain two language-related components. For each sentence, the model must tell us what parts of the world the sentence is about, and in which worlds it is true.

Our answer to the meaning question appeals to two relations: sufficiency and production. We propose that the meaning of *cause* and *because* is a blend of these relations: *C cause E* and *E because C* are true just in case *C* is true, and *C* is sufficient to produce *E* but *C*'s negation is not.

cause = *difference-making* + *sufficiency* + *production*



INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION