Mind as Theory Engine:
Causation, Explanation and Time

By
Michael D. Pacer

A Dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Psychology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Thomas L. Griffiths, Co-chair
Professor Tania Lombrozo, Co-chair
Professor Alison Gopnik
Professor John Campbell

Fall 2016

Abstract

Mind as Theory Engine: Causation, Explanation and Time

by

Michael D. Pacer

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Thomas L. Griffiths, Co-Chair

Professor Tania Lombrozo, Co-Chair

Humans build theories out of the data we observe, and out of those theories arise wonders. The most powerful theories are causal theories, which organise data into actionable structures. Causal theories make explicit claims about the structure of the world: what entities and processes exist in it, which of these relate to one another and in what form those relations consist. We can use causal theories to induce new generalisations about the world (in the form of particular models or other causal theories) and to explain particular occurrences. This allows rapidly disseminating causal information throughout our cognitive communities. Causal theories and the explanations derived from them guide decisions we make, including where and when to look for more data, completing the cycle.

Causal theories play a ubiquitous and potent role in everyday life, in formal pursuit of them in the sciences, and through their applications in medicine, technology and industry. Given this, the rarity of analyses that attempt to characterise causal theories and their uses in general, computational terms is surprising. Only in recent years has there been a substantial refinement of our models of causal induction due to work by computational cognitive scientists — the interdisciplinary tradition out of which which this dissertation originates. And even so, many issues related to causal theories have been left unattended; three features in particular merit much greater attention from a computational perspective: generating and evaluating explanation, the role of simplicity in explanation choice, and continuous-time causal induction. I aim to redress this situation with this dissertation.

In Chapter 0, I introduce the primary paradigms from computational cognitive science – computational level analysis and rational analysis – that govern my research. In Chapter 1, I study formal theories of causal explanation in Bayesian networks by comparing the explanations the generate and evaluate to human judgements about the same systems. No one model of causal explanation captures the pattern of human judgements, though the intuitive hypothesis, that the most probable a posteriori explanation is the best performs worst of the models evaluated. I conclude that the premise of finding model for all of human causal explanation (even in this limited domain) is flawed; the research programme should be refined to consider

the features of formal models and how well they capture our explanatory practices as they vary between individuals and circumstances. One feature not expressed in these models explicitly but that has been shown to matter for human explanation is simplicity. Chapter 2 considers the problem of simplicity in human causal explanation choice in a series of four experiments. I study *what* makes an explanation simple (whether it is the number of causes invoked in or the number of assumptions made by an explanation), *how* simplicity concerns are traded off against data-fit, *which* cognitive consequences arise from choosing simpler explanations when the data does not fit, and *why* people prefer simpler explanations.

In Chapter 3, I change the focus from studying causal explanation to causal induction — in particular, I develop a framework for continuous time causal theories (CTCTs). A CTCT defines a generative probabilistic framework for other generative probabilistic models of causal systems, where the data in those systems expressed in terms of continuous time. Chapter 3 is the most interdisciplinary piece of my dissertation, accordingly it begins by reviewing a number of topics: the history of theories of causal induction within philosophy, statistics and medicine; empirical work on causal induction in cognitive science, focusing on issues related to causal induction with temporal data; conceptual issues surrounding the formal definition of time, data, and causal models; and probabilistic graphical models, causal theories, and stochastic processes. I then introduce the desiderata for the CTCT framework and how those criteria are met. I then demonstrate the power of CTCTs by using them to analyse five sets of experiments (some new and some derived from the literature) on human causal induction with temporal data. Bookending each experiment and the model applied to it is are case from medical history that illustrate a real-world instance of the variety of problem being solved in the section; the opening discussion describes the case and why it fits the problem structure of the model used to analyse the experimental results and the closing discussion illustrates aspects of the case omitted from the initial discussion that complicate the model and fit better with the model introduced in the next section. Then, I discuss ways to incorporate other advances in probabilistic programming, generative theories and stochastic processes into the CTCT framework, identify potential applications with specific focus on mechanisms and feedback loops, and conclude by analysing the centrality of temporal information in the study of the mind more generally.

Excepting the supporting appendices and bibliography that end the dissertation, I conclude in two parts. First, in Chapter 4, I analyse issues at the intersection of three of the main themes of my work: namely, (causal) explanation, (causal) induction and time. This proceeds by examining these topics first in pairs and then as a whole. Following that, is Chapter 5, an epilogue that clarifies the interpretations and intended meanings of the "Mind as Theory Engine" metaphor as it applies to human cognition.

In memory of David J. Pacer.

Contents

Listing of figures

Listing of tables

Preface

> *Here was a first principle not formally recognized by scientific methodologists: When you run onto something interesting, drop everything else and study it.*
>
> Skinner [9]

The work that follows – especially that in Chapter 3 – is a product of following Skinner's [9] principle.

When I first began studying explanation, I knew it was a fraught subject studied by many before me and there was a great deal of foundational research to build my work off of. Accordingly, I set out to read much of this work, and learned a great deal from it. Tania Lombrozo was phenomenal in her ability to assist in this manner. I thought it would be the same with causation, induction, and time; surprisingly, the reality could not have been further from the truth.

One might have thought the situation to be reversed; time and causality have been perennial subjects studied by many more people across many more fields. And that is true, many people have studied time and causality before, even time and causality at the same time. The problem was not that I had no guidance — Tom Griffiths was as remarkable in his ability to find resources and direct me toward related topics as he was insistent in remarking I needed to stop finding new aspects of the problem and additional topics, but to instead focus on completing the work I had begun. As hinted at by that remark, the problem was not that no one had studied these, but that – considered as a whole – the research that had been done on these topics was so broad, deep and diverse that the resource I was looking for (i.e., one that would provide a comprehensive foundation off of which I could build my work) simply did not exist. No one had even surveyed the landscape on which the foundation could be built.

It is when faced with such a problem that Skinner's [9] advice becomes problematic — when you are faced with a labyrinthine literature that offers something interesting and new at literally every turn, you end up "dropping everything" as rapidly as those "everythings" accrue. With no map to work off of, you will end up lost and aimless.

Fortunately, I entered the maze with my own version of Ariadne's thread: pen and paper. Before graduate school, you would rarely find me without a notebook and writing utensil of some kind; now such instances are nigh impossible. So, though I may not have had a map going in, once I realised the challenge I faced, I began my bit of intellectual cartography and thereby mapped my way out of the problem.

What you find in Chapter 3 is not a complete survey of all that I studied and read — far from it. Though what is represented is a only a small fraction of the total area that merits exploration, this area though has a most useful property for beginning a foundational project. It is a map of those subtopics that were most closely linked to the home topic (the study of causal induction, time and human cognition in the cognitive sciences) that were able to be tightly integrated with one another (were conceptually closely related), while also not having yet been integrated with the existing literature. It may be that this can be a criterion for determining importance in a explore-exploit situation where the number of available states is unknown at the time of exploration but where a relative relevance function is available by which you can compare the quality of nearby states with joint reference to your original aim and whatever the state's current value is.

In a sense, this criterion was the inevitable conjunction of Skinner's methodological advice, his theories of reinforcement, and his comment[10], "Education is what survives when what has been learned has been forgotten." That is, the paths I traversed the most are the ones I travelled the most. And as a result the map is of those areas that were most simultaneously most interesting (and in a sense novel compared to other readings) while also being interestingly accessible from as many other interesting points as is possible. As long as we have some memory of where we have gone before – provided in my case by notebooks – Skinner's strategy proves to be an excellent way to identify not only interesting areas of research/history/mathematics, but those areas that are most closely connected to other interesting areas that otherwise appear to be quite remote (and therefore interesting) to each other.

As a cognitive scientist, I find it amusing how frequently and perfectly B.F. Skinner, the most adamant anti-cognitivist in recent memory, seems to have taught the lessons I have learned in the vein of arguing against the possibility of the topic of my studies being part of a scientific discipline.

The greatest challenge in making the map were the autodidactic paths that journeyed through more technical topics. There is always much background work leading up to advanced work in technical fields. Accordingly, the cutting-edge research in technical fields is written for those who already have that background. But, what is efficiently conveyed to practitioners of the field omits thorough introductory materials, and thus is incomprehensible to all but the most dedicated novices. This would pose little issue, except that the advanced research proved relevant while the basic research did not. This meant I could not expect others to know how to follow in my footsteps without my having laid a complete trail. Nonetheless, it was clear early on that the foundation I was building would need to be constructed out of the sturdy materials and techniques only available in technical literature. Only by using these techniques would I have the mathematical and computational precision to model the real problem that human minds actually fact with the veridicality that I sought. Learning something well enough to use it is fairly easy; learning something well enough to teach it to others and to extend it is challenging; but most challenging is to learn something well enough to teach *others* how to build off of that knowledge is the challenge. But it was the last task that that was before me as the builder of a new foundation that I hope will be the bedrock to much more work than any one person can

achieve in their lifetime. Given the reactions I have received already in response to Chapter 3, I feel confident in my decision; the effort has paid off.

I encourage anyone who wishes to take a Skinner-search approach to their work to be emboldened when they run into challenges from technical topics. To have even learned that you need to understand a topic suggests that you have a capable mind — one capable enough to learn the topic in question. And, I have found that the harder it is to gain knowledge the more profitable it has been. You share a background and research with many of your colleagues, accordingly those who do not even know that a topic exists that is deeply relevant to their inquiries will benefit from the effort you put into converting the esoteric knowledge into terms closer to those they are familiar with. Better yet, if you can illustrate why it is important to their particular interests, you will truly be a boon to those around you. In my case, the extensive study of continuous-time stochastic processes and their formal foundations allowed me to share insights unavailable to my colleagues, despite the pervasive relevance of continuity to computational and cognitive problems.

This reveals what I suppose was the heuristic that (once recognised in afterthought) guided my choice of which areas to map out: map with greatest care those areas that are the most treacherous and likely to discourage the easy traveller. This is part of the other reason for Chapter 3's length — I wanted to provide enough of a background for anyone who wished to follow the routes I have laid out that they would be able to do so while needing to consult other guides as minimally as possible. Yes, the goal of the chapter is to provide a formal foundation out of which we can reconstruct extant causal modelling research while also accounting for cases and phenomena that could not have thought by one who only had access to the classical approaches. Building off of such a foundation requires facility with topics that even the best educated cognitive scientists could not be expected to have in their tool-belt. I wanted to ensure that all those so enticed by the possibility my research programme offers would have the resources needed to do make those possibilities realities.

I know that few people read dissertations at all, and that fewer few read them in their entirety. Nonetheless, I did and do and I know that there are others who have and will.

For those who do come across this, my hope is twofold. First, that they will find their way made easier by the path that I have laid before them. And second, that it will introduce them to ideas interesting enough that they stop reading it and pursue that more interesting thing — at least, until something else reminds them of this dissertation as an interesting place to look, at which point their Skinner search leads them back to this place to continue the journey (renewing it from where they had left off or from whichever part happens to be most interesting at the time).

Acknowledgments

Frank Keil, were it not for your support of my early research career, I likely would not have written this document. I learned a great deal from working with you, and I am proud to think of it as a long-range consequence of our work together at Yale. Thank you.

Kevin Uttich, Nick Gwynne, Dylan Murray, Carly Giffen, Sara Gottleib, Nadya Vasilyeva, Azzu Ruggeri, and Daniel Wilkenfield; Naomi Feldman, Saiwing Yeung, Anna Rafferty, Jing Xu, Liz Bonawitz, Daphna Buchsbaum, Josh Peterson, David Bourgin, Rachel Jensen, Daniel Reichman, Aida Nematzadeh, Nori Jacoby, Alex Paxton and Dawn Chen: You all have been excellent colleagues who have taught me a great deal. Thank you.

Lisa Clark, Tiffany Wu, Aria Pakatchi, Ginny McGinnis, Kevin Serrano, and Joy Hillregel, mentoring you all was a privilege, I learned a great deal. Thank you.

Rafa Kern, your influence on me over the years that I built the body of work represented here is strange and profound. I think neither of us would have it any other way. Thank you.

Joseph Williams and Brian Edwards, we each have travelled far together. The journey has been interesting and delightful. You excel in your companionship; others are stronger when they are around you. Thank you.

Joe Austerweil and Karen Schloss, you will forever have a warm place in my heart from their kindness toward me from when I first arrived at Berkeley to today and beyond. My maths would be much weaker, and my posters much less pretty were it not for your influence. Thank you.

Brian Waismeyer and Anna Waismeyer, we have shared many great times and fanciful discussions, while those same discussions have led me to rethink and rework aspects of my research programme. Rare is the conjunction of the two. Thank you.

Kevin Canini, your acknowledgements inspired me to write these. Yours made me proud to be your friend and to have supported you around dissertation time in any way. Thank you.

Caren Walker, your accomplishments speak for themselves and it has been a pleasure to see the world recognise the insight you bring to everything you think about. I can hardly wait to learn and think about the next thing you identify amongst the many facets of learning by thinking that you will explore. Thank you.

Stephanie Denison, I still hope that we will find time to study time together. Because of our discussions, I have thought much more deeply about the role of time in the study of infant cognition both as a topic and as a methodological tool. My work has benefitted greatly from it. Thank you.

John Schindel, you have helped in ways no one else could in a number of occurrences. I think it is safe to say that I would not have graduated were it not for your help. Thank you.

Sophie Bridgers, Dave Schwantes, Dillon Plunkett, Ellie Kon you have all made my life more pleasant and easy. Thank you.

Chris Holdgraf, you have made my life harder. I am the better for it. Thank you.

Claire Oldfield, keep doing what you do. You inspire the rest of us. Thank you.

Lauren Harris, I cannot help but smile when I think of you. We both have had major constraints on our time, fortunately the existence of this sentence suggests that mine is soon to be alleviated. I look forward to being able to spend time together again, it is always a delight. Thank you.

Stéfan van der Walt, you never cease to convince me that more can be done and that greatness can be achieved with diligence. You have always supported my work in a unique way; you make me proud to be studying the things I study in the way that I study them. Thank you.

Nathaniel Smith, though the final form of our work on improving the efficiency of my algorithm only is implicitly referred to in Appendix C, the act of going through it made many other pieces fall into place. As a result, our work made writing the time chapter much more straightforward when it came time to do so. Thank you.

Matthias Bussonier, I am fortunate that your reservoir of patience seems to have no finite measure. I am genuinely a different person for your tutelage and friendship. Thank you.

Fernando Perez, Jaime Whitacre, Stacey Dorton, your support and aid in transitioning from the project of writing this dissertation to joining Project Jupyter full-time. Thank you.

Sarah Wellen, you played a key role in helping me move forward when I ran into the limitations of the written record of data. You gave me faith in my work. Thank you.

Neil Bramley, you are unique in your ability to challenge my way of thinking about problems. In doing so you have push me to work harder and develop better explanations and better founded theories. My work is much better for having known you. Thank you.

William Greville, your contribution of the stimuli from Greville and Buehner [2] was vital to the initial stages of developing the projects that led to Chapter 3. Thank you.

David Lagnado, were it not for your generosity in sharing your experimental stimuli and forthrightness in discussing your results, Chapter 3 would not be have been possible. I have enjoyed our virtual conversations and I hope that we will manage to have them in person some day. Even more so, I hope that we will manage to work together on any of the many topics that interest us both. Thank you.

Alex Carstensen and Terry Regier, though we worked together on only one project, the process of doing so was immensely informative. I have been a better collaborator and scientist as a result of our past and present interactions. Thank you.

Stephan Meylan, you challenge me in ways others either do not or dare not, I do not know which (though if it is the latter, I wish more would). You oscillate between the realistic, the idealistic and the theoretical with a rare fluidity that I wish I could emulate. And you do all of this with poise. I do not know how you manage all of this, but regardless I appreciate your being you and its effect on me and those around me. Thank you.

Falk Lieder, I have never met someone so capable of carefully thinking things through as you are. I and the world have much to learn from you; fortunately you are willing to teach us

a great deal. I just hope that the world will find a way to offer you even a tenth as much as you offer it. Thank you.

Josh Abbott, we made it. It was never in doubt, but it has been a long journey. I am grateful for having shared this road with you, even if we never did manage to be in the office at the same time. Thank you.

Thom Morgan, I am a better scientist and a better thinker in general for the time we spent in conversation. I never knew how much my thoughts on evolutionary theory were in need of refinement until speaking with you — and this is after having read many books on the philosophy of biology! Reading even the first part of West-Eberhard [11] together gave me a unique, and powerful appreciation for many aspects of the world and our studies of it that I simply had not had access to. I already miss our coffee breaks and wish I had allowed writing document to interfere less with our taking them. Thank you.

Andrew Whalen, your dogged ability to accomplish what you set your mind to is endlessly impressive. Your competency at such a diverse set of skills, combined with your determination, means that rarely will anyone need to worry for you; you ensure that you always do well. That is a gift you give to those who care about you that we are all grateful for. I look forward to seeing where your goals next take you. Thank you.

Luke Maurits, instead of writing you a letter, I wrote you this acknowledgement. Unlike a letter, odds are good that you may never read this, and sure that if they did someone else has read it before it reaches you. You awaken me to the limitations of media. You are one of the few people that can find things that I still will initially think are boring (e.g., vintage Chinese watches) and convince me that they were deeply fascinating once I knew enough to notice how and why. Please, never stop tinkering. Thank you.

Chris Lucas, I genuinely think that you and I may have spoken more about the topics reported here (and so many others) than anyone else I know. And, of course, this mostly occurred while we were 8 time zones apart. You never cease to astound me with your technical insight, in terms of both depth and diversity. At the same time, you also never let that expertise stand in the way of explaining something in terms I can understand. Thank you.

Avi Press, Dan Scott and Peter Chen each of you contributed in major ways to one or more of the projects described here. Working with you has been extremely rewarding. And while working together taught me a great deal (per se), you each forced me to think through what it was to be a good mentor and to help foster a career. I am proud of what we accomplished together, and look forward to seeing all that you continue to create. Thank you.

Jordan Suchow, I literally could not have written this dissertation were it not for you*. You have encouraged me to grow and accomplish things like no one else ever has. I do not know if I will find someone else who shares with me the capability and tendency to hold in mind mutually incompatible perspectives on the same topic. I am sure that if I do, it will be unlikely

* This dissertation was written using 'Dissertate' a templating framework for defining dissertation templates for many different schools' formatting requirements. For this, I built out the UC Berkeley template.

that they will not share an interest in so many topics that I hold dear, which has allowed what is possibly the most productive conversational relationship that I have ever observed(let alone participated in). We have already built great things together, and I anticipate that we will continue that trend. Thank you.

Jess Hamrick, I don't know how to convey to you the degree to which you have deeply affected my life. There is likely not a page in this left unaffected by the knowledge you imparted, the questions you posed, and the counsel you offered. Many people are able to produce amazements that give me hope for what can be accomplished by people who put their minds to a task. You do that but go further — you give me hope for what *I* can accomplish. I cannot think of a tale of my accomplishments in the near future that does not resound with the echoes of your influence. May the effects you bring elsewhere in the world be as monumental and as monumentally positive as the effects you have brought to my life; this is a wish I have no doubt will be granted. Thank you.

Gregory & Cassie, you've always been there to support me; I hope some day my work can play a role in making your lives better. I love you. Thank you.

Mom, think of my words, programs and deeds as your grandchildren; they are the means by which any legacy I have will persist and I am proud of them. I am confident you will be too. I would not have been able to accomplish any of what I have were it not for the sacrifices you and Dad made while I was growing up. I love you. Thank you.

E & Callisto. You helped *us* survive **me**. And – for that – I could neither love nor thank you enough.

My committee

John Campbell, thank you for witnessing my academic journey. I will never forget your reminder to reduce the focus on the difficult unknown details and to put more effort toward shedding light on the core issues that would benefit from greater attention. I hope that the lesson is reflected throughout the dissertation. Thank you.

Alison Gopnik, your work has always inspired me. What you have decided to talk about and the manner with which you talk about it has dramatically affected my own subjects of study. I especially appreciate your manner of accepting critiques and your ability to nimbly respond and cut to the heart of the source of that critique. I still consider your suggestion about writing a review article about the history and philosophy of science and its relation to computational models used in the computational cognitive sciences. Now that I will be done with my dissertation, it may be time to revisit that project. Thank you.

Tania Lombrozo, when I first joined your lab, I was in awe of you and all that you accomplished. That awe persists, and it has been supplemented by a deep appreciation for the speed, poignancy and clarity of your mind about an astounding array of issues. Through discussion with you and reading – you enabled me to take questions that I had thought had boring, easy answers and completely reshuffle my perspective. In the end, because of your questions and

suggestions, I recognised that there was unseen complexity and subtlety to these and many issues that, in my haste, I had either dismissed as details or failed to consider at all. You have made me a better thinker, and for that I can never thank you enough. You helped me learn to redirect my tendency to branch off and discover new topics into a argument honing mechanism, by which I approach different aspects of the same argument and make it stronger and more incisive with each pass. Thank you.

Tom Griffiths, you have guided me from someone who did not understand what a causal Bayes net really was and turned me into a scholar capable of inventing a formal framework that reinvents what it is to have a network in time. As time goes on, my tendency to look into everything that is even somewhat interesting has always made me feel like cognitive science is only one of many things that interests me and other endeavours (e.g., machine learning) begin to look just as appealing. But then, every time you give a talk, my desire to study cognitive science is rekindled and I see other topics through the lens of what they can reveal about the human mind. At the same time your ability to impart technical skills and insight by talking through mathematics, algorithms and computations is unmatched. You introduced me to entire fields of study from which, were it not for your guidance and assurance, I would have been rapidly discouraged and fled. You have tempered my mind and taught me to push through the feelings of incompetence that arise when faced with the unfamiliar and complicated. As a result, you taught me to be able to teach myself not only topics that come easily, but those which seem most impossible and out-of-reach. Also, my `uber.bib` product by descent and modification of yours, that saved a lot of effort over the years. Thank you.

Tania and Tom, in addition to your individual prowess as advisors, you have worked amazingly well as a team, and I appreciate the challenge that that can pose. You have helped me more blunders than I wish to remember; but, even in doing so, you were always positive and supportive. True *focus* may still elude me, but with your help I have learned to make a concerted effort and thereby see progress. Between honing and tempering, my mind is now a much more formidable weapon in the arsenal of computational cognitive science. But even more, I am not only a better academic for having worked with you, but a better person. Thank you, both.

Original collaborators on material included within

Much of the described in the following pages derives from work conducted in concert with others. The work in Chapter 1 is based off of Pacer et al.[12] in which Joseph Williams, Peter (Xi) Chen, Tania Lombrozo and Tom Griffiths. The work in Chapter 2 is work joint with Tania Lombrozo and is derived from a paper that is currently under review. I was assisted in the coding explanation and simplicity results by Peter Chen, Avi Press and Daniel Scott. The work in Chapter 3 is joint with Tom Griffiths and is amalgamated partly from previous conference publications Pacer and Griffiths[13],[14] as well as a substantial amount of unpublished work that first appears in this dissertation. Some of the experiments and analyses described in the time

chapter feature contributions from Avi Press and (especially) Dan Scott. Additionally, data from Greville and Buehner [2], Lagnado and Sloman [6], and Lagnado and Speekenbrink [7] were provided for analysis in Chapter 3, for that I am exceptionally grateful.

Funding Sources

The work reported within this dissertation could not have been undertaken without the financial support from a variety of sources. My efforts in building the material that make up this dissertation was supported by a Berkeley Fellowship and a National Science Defense and Engineering Graduate Fellowship. The individual chapters were supported under a variety of specific grants. The work in Chapter 1 was supported by grant DRL-1056712 from the National Science Foundation to Tania Lombrozo, and grant IIS-0845410 from the National Science Foundation to Thomas Griffiths. work in Chapter 2 was supported by grant DRL-1056712 from the National Science Foundation awarded to Tania Lombrozo. The work in Chapter 3 was supported by grants FA9550-10-1-0232 and FA9550-3-1-0170 from the Air Force Office of Scientific Research awarded to Thomas Griffiths. Much of the final writing of this dissertation was supported by the Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science Grant awarded by the Helmsley Trust, the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation to Project Jupyter. The Berkeley Institute for Data Science has provided an excellent place to work on this and related work.

0 Introduction

Minds exist. They affect the world, and those effects demonstrate their existence. They are as causally potent as any government, malady, law, corporation, or operating system. These systems affect the world in accordance with their representations of the world.

Mental representations can have rich structure, notably this includes causal theories. Causal theories describe the way minds encode the events, states and processes that make up the world. In turn these structured mental representations inform interpretations and guide actions. If it can be managed, research on the structure and form of those causal theories will be a key contribution to the sciences of the mind.

However, attempting to pursue this activity and to scientifically understand the structure and form of causal theories embedded in the mind will require great precision and care in the manner of investigation. Formal precision is needed to ensure that in the effort we are as clear as possible in our meaning — that others will be able to comprehend exactly which claims we are making. Care in experimental design is needed to ensure that the investigation is targeted at identifying the consequences of the causal theories themselves.

The work that follows is an effort in this endeavour. In it, I will investigate the match between formal theories of causal explanation and human explanation generation and evaluation, the role of simplicity in explanation choice, and causal induction using a variety of kinds of continuous-time data. I will rely on formal work drawn from a number of fields, this in-

cludes work on formal accounts of explanation [15,16,17,18,19,20,21], formal accounts of simplicity [22,23,24,25,26,27], causal Bayesian networks [28,29] and causal theories [30,1]. I will rely on the empirical methods of experimental psychology to assess:

1. which explanations people generate(section 1.4) and evaluate well(section 1.5) when given probabilistic descriptions of the scenarios in question.

2. what kind of simplicity informs human explanation choice(section 2.2), how this notion of simplicity interacts with observed data (section 2.3), what the consequences of choosing simpler explanations are (section 2.4), and in what cases that simplicity has stronger effects (section 2.5).

3. how people infer the form and structure of causal systems in the case of continuous time data such as rates (section 3.9), tables (section 3.10), trials (section 3.12, and real-time displays (section 3.11 and section 3.13).

All of this work touches on issues drawn from cognitive psychology, computer science, and philosophy; it can only be said to be encapsulated within the domain of cognitive science*. Each chapter discusses a collection of work, for which I give a self-contained introduction to the topics that directly relate to that work. I will rely on those introductions to address the specific academic literature that drives and informs that work. For the remainder of the introduction I will introduce issues that are of a more general scope and touch on the larger themes around which my work is organised.

0.1 Levels and styles of analysis

The work that follows is in the tradition of computational-level analyses (as described by Marr [31]) and rational analysis (as described by Anderson [32]).

0.1.1 Computational-level analyses and the ubiquity of representation

Computational-level analyses allow us to connect formally precise models of problems to human behavior in solving an analogous problem. In Marr's [31], p. 24 original words,

* and even then only if one allows history of medicine to be included in cognitive science

> the abstract computational theory of the device [characterises the performance of the device] as the mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriate and adequacy for the task at hand is demonstrated.

More generally, it allows us to describe what it is that an information processing system "aims" to accomplish[†]; that is, what the problem is that motivates it. This kind of problem definition well define both what kinds of solutions count as valid (based on the available computational actions) and in describing why it does the things it does it can provide an ordering on the quality of the available solutions (in terms of how well they accomplish its aim). Computational-level analyses are contrasted against algorithmic-level analyses (how the solution to a problem is represented in terms of the available algorithmic actions) and implementation-level analyses (which explains how a solution to a problem is implemented in terms of physical primitives and their interactions).

This description differs somewhat from the standard given for distinguishing between these levels. Algorithmic-level analysis is sometimes called the "representational-level" analysis, but this is a confusing naming convention, so I will not use it. To put it briefly, every level of analysis will need to imbue its formal structure with some representational primitives in order to be capable of being expressed as a model. Euclidean geometry in this sense is a representation of space. The notion of registers, virtual logic gates and binary operations are representational primitives of the implementation of most modern computing systems; but so are the notions of a variable, data structures and pointers that describe the primitive operations and higher-order representations available as described in the beginning of algorithms books. But these representations need to work well with one another in order for the levels to make coherent statements when considered together. It may never be the case that in practice one level is reduced to operations and representations at another, but for the programme to function that needs to at least be a possibility. And for that even to be a possibility, all the levels must have operations and representations of some kind, otherwise their formal equivalence (which is a representation in and of itself) would be impossible to express.

The classic metaphor for distinguishing between these levels is three different ways of investigating flight, particularly bird flight. At the implementational level this would consist of

[†] This is not to suggest that information processing systems aim in the sense that we attribute to agents except to the extent that those agents themselves can be described as computational systems with feedback loop based behaviour that allows some degree of equifinality. This is discussed at greater length in subsection 3.14.5.

the study of bird feathers, their composition, shape and perhaps even there relative location on birds' bodies. At the algorithmic level this would involve studying the sequences of actions by which birds actually fly (cataloguing the different way bone and musculature structures interact with each other during the course of flight). At the computational level the explanation would describe the principles of aerodynamics in an abstract mathematical form. Marr [31] introduced these levels in part because of what he saw as an overemphasis on the lower two levels (where neuroscientists studied the implementational level and experimental psychologists studied the algorithmic level) and a relative ignorance of the computational level. As he pointed out, early efforts at human flight attempted to mirror the motions that birds engaged in, and accordingly fared poorly. The recognition of the problem as being one of aerodynamics freed people from the metaphor of bird flight and allowed the creation of kinds of machines that fit the aerodynamic requirements but which resembled no bird that has ever existed.

In the case of human vision, Marr [31] identifies producing object representations of the world as the computational level analysis of the visual system. The mechanistic level of analysis identifies the neural mechanisms that compose the physical implementation of the solution to the problem described at the computational level. The Hodgkin and Huxley [33] model of the neuron as a circuit diagram is a formally precise analysis at the mechanistic level of how action potentials propagate. The algorithmic processes used by the visual system to translate the input from the eyes to the resulting signal that meets the goal defined by the computational-level analysis, often under constraints imposed by the particular implementation (e.g., a limited number of algorithmic operations could be allowed on the basis of the speed of the process combined with our knowledge about the processing speed of individual neural signals). Stereopsis (how you combine the signal from two separate channels of information, your eyes) would be included in the algorithmic level of analysis. Stereopsis would not sit at the computational-level, but the computational-level would determine the quality of the algorithmic solution. Stereopsis would also not sit at the implementation-level as it is a phenomenon that is multiply realisable in neural architectures and systems organised so as to be extremely similar at the level of neurons could vary widely in their ability to implement a stereoptic algorithm.

My work sits at the computational level of analysis and explicitly does not focus on the other levels of analysis. Given the lack of neuroscientific research, it should be straightforward to see that my work is not at the implementational level. It may be less straightforward to see that my work does not sit at (and should not be interpreted at) the algorithmic level of analysis. I do not provide arguments regarding the occurrence (or non-occurrence) of particular mental

procedures that operate over "mental" primitives. I ask people to generate, evaluate and choose explanations, but those tasks are formally defined using mathematics external to however it is that the problems are to be represented in the mind. I do investigate the consequences of the solution of computational problems for other processes (such as memory, see section 2.4), but I do so in a way that eschews the algorithmic level. I do not consider how memory is represented at any more fine-grained a level than someone's ability to exactly reproduce summary statistics that are exactly known because they were generated as experimental stimuli. Statements that a particular decision modulates people's responses is consistent with many models of human memory.

Put in a slightly different way, my concern is how any computational system (implemented however on whatever algorithmic architecture) would solve the problems of causal explanation and causal induction. From this perspective there is no a priori privilege given to the human mind, but humanity establishes a standard that no other system has matched in its success.

0.1.2 Rational analysis and quasi-optimality

A computational level analysis provides a framework for expressing formal structure of a problem. It even expresses what it would be to solve the problem well. However, it does not require that (to be applicable) that any particular system will solve any particular problem well. One of our motivations as to why human causal inference is an interesting problem is precisely people's facility with it. It would seem that we would like a methodology for accounting for this success that extends beyond merely what defining success would be. For that we turn to rational analysis[32].

Anderson[32] poses rational analysis as a way to realise incorporating his "General Principle of Rationality" [‡] into a systematic method for research in cognitive science. Anderson[32] describes rational analysis as:

1. Precisely specify what are the goals of the cognitive system.
2. Develop a formal model of the environment to which the system is adapted (almost certainly less structured than the standard experimental situation)
3. Make the minimal assumptions about computational limitations.…

[‡] From Anderson[32], "*General Principle of Rationality*: The cognitive system operates at all times to optimise the adaptation of behaviour of the organism".

4. Derive the optimal behavioural function given items 1 through 3.

5. Examine the empirical literature to see if the predictions of the behavioural function are confirmed.

6. If the predictions are off, iterate.…

In the sense of following exactly those steps, my work is not an instance of rational analysis. But in other senses it is.

0.2 Rational, computational analyses of explanation and causal induction

Throughout the following work, I aim not only to define the abstract structure of the problems but to identify different ways in which, given a class of problem structures, one might solve it optimally. This is why, in Chapter 1, I give justifications in the vein of a computational-level and rational analysis for multiple formal models for causal explanation. These models are framed in the language of exactly defined Bayesian networks. Models of these sort presume that the problem that explanation aims to solve is one where you know the relevant causal system as it applies across cases, and that you have observed some of the variables values. In that case, to produce an explanation is to determine which variable settings in the network would best explain a subset of those values.

The intuitive answer as to what would "best explain a subset of those values" to many seemed to be the "obvious" original answer – the one taken for granted to be *the* answer of what it is to provide an explanation in a Bayesian network in Pearl [15]. Namely, the best explanation was that setting of the unknown variables that had the highest a posteriori probability (conditional on your knowledge of the network's parameterisation and all the observed variables). But, other researchers found other ways of defining what it what needed to be included in the explanation (e.g., perhaps not all the unobserved variables [16]), what needed to be explained (e.g., not all the observed variables [19]), what criterion should be optimised (e.g., the generalised Bayes factor [21]), whether the criterion for generating explanations should be the same as that for evaluating them (e.g., evaluate based on posterior probability but generate based on information theoretic measures [18]) or what could count as an explanation (e.g., perhaps some of the observed variables [19]). This might be uninteresting variation were it not for the fact that these

different accounts made vastly different predictions about which explanations were to be preferred in the context of particular model problems. With so many available accounts this offers an excellent opportunity for a study of human performance on just those analogous problems. The model performances can be expected to differ and so the problems will be discriminative (if any of the models are successful at all at predicting human behaviour). At the least, doing so gives a comparative account of the different models and examples of how to built experiments that induce people to provide analogous judgements to the models themselves. And even if none of the models succeeds, if people provide the same answers in situations where the models differ (or vice versa), this should give insight as to the degree people attend to the different criteria that feed into the building of a model.

Even so, every one of these models of the problem of explanation clearly fails as a rational, computational-level analysis of all human causal explanation. At the least, not all problems analysable by human causal explanation may be described in terms of perfectly known Bayesian networks about which we have perfect observations of a single trial. However, that does not mean that we cannot learn a great deal by investigating to what degree people, when put in that situation, generate and evaluate explanations in accordance with the predictions of the various formal models. Remember that rational analysis treats the environment as given, and there may be some cases in which people need to solve analogous problems and this will give us insight into their approaches to those problems (if any of the models succeed at capturing human behaviour whatsoever). Accordingly, we had to translate these networks into stimuli capable of being understood by people who we could not assume were acquainted with the intricacies of Bayesian networks. Instead we encoded the networks in terms of a causal scenario that mirrored the formal structure and gave data consistent with the parameter estimates that defined the networks links. People were able to use these stimuli to generate and evaluate explanations for particular observations, and we had them do so in ways that we could directly compare to the predictions of the various models.

The models themselves varied widely in their predictions over the stimuli. The networks were drawn from the literature in which the explanation models were introduced for the purposes of displaying differences between the predictions from different models. Usually those differences were framed by the newly proposed model as shortcomings in the models that the new one was trying to supplant. So it should be unsurprising that they disagree about which variable settings make for the best explanation. But that says little about which of these models is *best*. We operationalise that in terms of which of the models best matches human explanation generation and evaluation.

More generally, the problem of determining the "best" model will depend on exactly which of the problems a system was optimised for. This holds for both cognitive and computational systems. Depending on the actions offered by a particular environment (e.g., whether one wanted a predictive or an interventional claim) the same optimal system could reasonably arrive at different conclusions. Indeed depending on the interpretation of the problem the same person could end up with different conclusions about what made for the best explanation.

From this work we learned that, regardless of how they interpreted, people did not attempt to maximise posterior probability when they generated and evaluated explanations. Models that focused exclusively on posterior probability poorly predicted people's responses. It seems that even if we have not been able to describe all of the cases of human causal explanation, we have at least ruled out the criterion that many (if not most) people think is "obviously" the correct answer when it comes to making these variable assignments. Ruling that out is a powerful advance.

None of the models of explanation that we studied in Chapter 1 incorporated an explicit simplicity metric in their judgements. But we know that, when it comes to causal explanations, humans prefer simpler explanations; it would make an already long document longer to list (twice) the many quotes from well-respected individuals who attest to the virtue of simpler explanations (see Chapter 2 if you wish to read them). Importantly, this holds both for lay and scientific explanations, and though the fact is agreed upon justification for why this preference exist is hardly discussed in the specific context of causal explanation. And, a formal account for why would be difficult to provide without first identifying the kind of why that would need to be.

And that is not to say that there are no formal accounts of why we might have a cognitive preference for simplicity. In fact, there are *many* formal accounts of simplicity that have a variety of consequences for human cognition, it just happens that they do not tend to apply well to the problem of causal explanation. Simplicity is argued for in terms of more efficient perceptual encoding (in the sense of a Shannon [34] encoding problem), minimal description lengths for cognitive programs, more efficient parameter and truth estimation, and effective allocation of probability/likelihood in defining the specificity of prior/likelihood functions. But, we found that nearly all of these simplicity justifications and their associated metrics fail to take into account the unique features of the problem of causal explanation. The types of accounts in the literature tend to rely on notions and metrics that do not use causal information explicitly, while this seems to be crucial for understanding a preference for simplicity in *causal* explanation (at least if causal explanation is to be distinctive in this regard at all). The one metric that

did relate was sourced from the psychology literature that we directly drew upon in generating our studies (for example, see Lombrozo [35]). However, part of the point of the work is to show that this metric fails in systematic ways at predicting people's behaviour in causal explanation choice tasks

We propose a new metric and established its ability to predict human behaviour better than the prior theory and with powerful further consequences for human cognition[§]. We found that when people rely on simplicity contrary to the evidence, there are apparently sub-optimal consequences for their memory for what they observed. However, in the vein of rational, computational analysis, we reanalyse the structure of the problem and establish cases in which this simplicity metric no longer seems to apply. From that perspective we see a different way to analyse the problem in causal explanation, especially causal explanation in the real-world. In that analysis a selective application of the simplicity metric could serve the goal of causal information compression. That is, that causal explanations that simplicity may allow optimally compressing our knowledge of the world such as to overrepresent as present those causes that play the role of the root of a causal system. One way of interpreting this is to describe the problem of dealing with the multifarious and profuse data in the world in the face of minimal cognitive limitations such as memory availability or processing time (making our account a less precise case of resource-rational analysis as in Griffiths et al. [36]). But, one could also reëxamine the problem as not being one of memory availability or processing time as traditionally considered, but time in the context of needing to actively respond to the world as the dynamic processes that make up the world continue.

If Chapter 2 describes a case in which we modify our model of the environment to accommodate a more complicated picture of the process, Chapter 3 is a paradigm showing just how much work can and must be done in order to appropriately reëncode the environment. One standard approach to causal induction in computational cognitive science, until recently has been a game played on contingency tables — counts of whether a cause and an effect did or did not coöcur over a number of trials. This approach can be seen to stretch back to Hume [37], Mill [38] and it makes assumptions that are thoroughly understandable at the time. That said, filling and computing with contingencies is simply not a problem that people face in everyday life. It is a creature spawned from well-intentioned attempts to be rigorous using the mathematical and conceptual frameworks available at the time. With reframing we should be able to go beyond this formal framework and discuss much more complex, rich phenomena such as those

[§] Technically, to predict human behaviour we needed to assume a decision model as well. Specifically we needed to show that people probability matched.

that occur in continuous time.

One of the other research traditions has considered phenomena that occur in continuous time. These are the research programmes based on conditioning (including both classical[39] and operant[40] varieties). Conditioning theories have similar Empiricist roots[37,38] to the theories of those wedded to contingency tables. Though the theories and models produced from the conditioning literature are capable of handling the continuous nature of experience with much greater verisimilitude, they fail to capture many of the most interesting features of human causal inference.

The world in which people exist is not just complex and continuous, but its complexity has a particular structure, and people are capable of inferring that structure. Not only are they capable of inferring that structure but they are capable of communicating about that structure to others. People can even induce causal relationships from stimuli that merely represent events as having occurred in time, without needing to reproduce that experience directly. They can do this in a variety of formats. The contingency theorists – themselves wedded to fairly simple inferential structures and the necessity of experiencing events in order to learn about the system producing those events – cannot account for people's ability to do this.

More generally though people don't merely infer causal relationships, but entire causal theories about the world from the data that they gather. They do so using a variety of kinds of information, combining different modalities and various conceptual structures into a common cognitive system for building and supporting causal theories. Those theories themselves can be communicated successfully from one person to the next, allowing the hard work done by previous generations to live on in the minds of future generations. This is the basis of scientific progress, it could not procede (let alone aggregate) if it were not for this ability. We have little knowledge of how this is possible, but that is in part because of the exclusion of the history of science and medicine from the cognitive sciences.

That human beings can accomplish something at all is a demonstration that the problem is one available to cognitive science as a source of inspiration or guidance. If we are looking for impressive cognitive feats that are well recorded, we could do much worse than to look to the history of medicine and science. Accordingly, interspersed with the technical apparatus and empirical methodologies in Chapter 3 are discussions and analogies drawn from cases in the history of medicine. The experiments and models were not directly about these cases, but it is intended that by juxtaposing them the similarities between the kinds of problems embodied in the cases and the kinds of problems addressed in the models and experiments in question will be mutually illuminating. This includes describing the formal structure of the informational

environment in which the various inferences were made; the models then make the intuitions about that formal structure and the manner of optimally solving it precise. The experiments show that these models capture not only exceptional cases of causal theoretic prowess, but inferential abilities that exist across a wide span of human minds.

1

Formal Models of Causal Explanation in Bayesian Networks: Evaluating formal models using human judgements*

1.1 Introduction

Representing statistical dependencies and causal relationships is important for supporting intelligent decision-making and action – be it executed by human or machine. Causal knowledge not only allows predictions about what will happen, but is also used in explanations for events that have already occurred. For example, a set of symptoms might be explained by appeal to a particular disease, or an electrical circuit failure by appeal to a set of faulty gates.

Previous work in machine learning has provided a range of models for what counts as an explanation in cases involving a known causal system and observed effects.These models differ in what they allow as potential explanations or 'hypotheses' as well as in the objective function they aim to maximize (for a review, see Lacave and Díez [41]). For example, one approach says hypotheses are settings for all unknown variables where you then choose the hypothesis that

* Some of the content in this chapter was originally published as Pacer et al. [12], co-authored with Joseph Williams, Xi Chen, Tania Lombrozo and Tom Griffiths.

maximizes a posteriori probability given observed data [15]; another allows hypotheses to be any non-empty variable setting and selects the hypothesis that maximizes the probability of observations under that hypothesis relative to every other hypothesis [42]. While these models differ in their formal properties, arguments for one model over another typically come down to which provides a better fit to researchers' intuitions about the best explanations in a given case.

In this chapter we evaluate four formal models of explanation by empirically investigating their fit to human judgements. Our aims are threefold. First, methods from cognitive psychology allow us to test how well competing models correspond to general human intuitions, rather than the intuitions of a small group of researchers. Second, by using human judgement as a constraint on formal models of explanation, we increase the odds of choosing an objective function with interesting properties for learning and inference. A growing literature in psychology and cognitive science suggests that generating and evaluating explanations plays a key role in learning and inference for both children and adults (for a review, see Lombrozo [43]), so effectively mimicking these effects of explanation in formal systems is a promising step towards closing the gap between human and machine performance on challenging inductive problems. Finally, formal models of explanation that successfully correspond to human judgement can contribute to the psychological study of explanation, as almost no formal models of explanation generation or evaluation have been proposed within the psychological sciences. In the vein of making these formal models more widely available to practicing psychological researchers, we have released an Explanation Engine on GitHub.†

We present two experiments in which we gave people information about a causal system and had them either generate explanations (Experiment 1) or evaluate explanations (Experiment 2). The causal systems can be formally defined by Bayesian networks and correspond to those used in prior work to differentiate among models of explanation [19,42]. Across two versions of two causal structures and across both experiments, we find that the Causal Explanation Tree [19] and Most Relevant Explanation [42] models provide better fits to human data than either Most Probable Explanation [15] or Explanation Tree models [18]. The results of our experiments identify strengths and shortcomings of these models, ultimately suggesting that human explanation is poorly characterized by models that emphasize only maximizing posterior probability.

† The Explanation Engine is written in `python` and, unfortunately, does not accord to the syntax specified for defining causal Bayesian networks in either Appendix B or Appendix C.

1.2 Bayesian networks

A Bayesian network provides a compact representation for the joint probability of a set of random variables, $\mathcal{X}$, which explicitly represents various conditional independence statements between variables in $\mathcal{X}$. We specify a directed acyclic graph with a node corresponding to each variable in $\mathcal{X}$. We say that each node $X \in \mathcal{X}$ has a set of "parent nodes" (Pa(X)), and that this gives us conditional probability distributions for every X given its parents $p(X|\text{Pa}(X))$. We assume that the full joint probability distribution can be specified this way, i.e., that $p(\mathcal{X}) = \prod_{X \in \mathcal{X}} p(X|\text{Pa}(X))$. This is equivalent to assuming that X is independent of all nondescendent variables given its parents, and allows us to use the structure of the graph to read off which conditional independence relations must hold between the variables [15].

Figure 1.1 shows an example of a Bayesian network specifying conditional probability distributions between random variables. The graph on the left (*Pearl*, named after Pearl [15]) represents whether a particular alien has a disease (D), whether that alien has a genetic risk factor for that disease (G), and whether or not the alien was vaccinated for the disease (V). The graph on the right (*Circuit*) can be interpreted as a circuit that always receives input and for which we can measure the output. A, B, C, and D are gates that, if functional, break the circuit, stopping the input from reaching the output. Each gate has an independent probability of failing and allowing current to cross through it. If the current can travel from the input to the output via any path made possible by a set of failed gates, then there will be output. These two examples hint at the richness of the Bayesian network formalism. We will continue to refer to these graphs throughout, which are the basis for our stimuli in Experiments 1 and 2, with the parameter values indicated in Figure 1.1.

1.2.1 Explanations in Bayesian networks

Suppose we observe the values for k of the variables in a graph, $\{O_1 = o_1, \ldots, O_k = o_k\}; \forall i O_i \in \mathcal{X}$. We may not wish to explain every observation, so let us call the variables we want to explain O_{exp}, with values o_{exp}. These values o_{exp} are the "target" of our explanation, or the *explanandum*, which is a subset of $\mathcal{O}$, the set of possible observation sets. We will refer to $\hat{O}$ as the set of variables that were observed and $\hat{o}$ as the observed values. Then $O_{\text{not-exp}} = o_{\text{not-exp}}$ are those variables that are observed and unexplained (or $O_{\text{not-exp}} \equiv \hat{O} \setminus O_{\text{exp}}$).

A candidate explanation (the *explanans*, or "hypothesis") is a set of variable assignments for some of the variables not in O_{exp}. We exclude O_{exp} to avoid circularity, though elements in

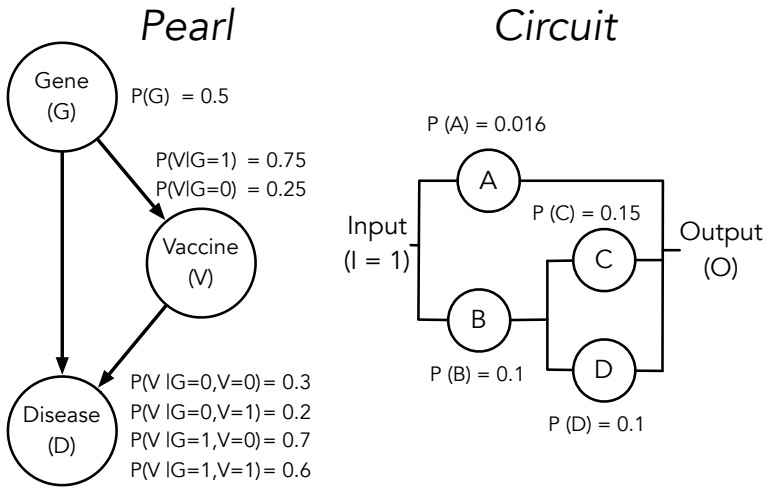


Figure 1.1: The *Pearl* and *Circuit* networks used in our experiments; as in Pearl[15], Nielsen et al.[19], Yuan and Lu[42] and Yuan et al.[21].

$\hat{O} = \hat{o}$ but not in O_{exp}(i.e., observed but unexplained variables) could be included. However, we should note that most models require that every observed variable be explained; formally $\hat{O} \equiv O_{\text{exp}}$. For the sake of clarity, a hypothesis (our term for potential explanans henceforth) will be represented by h, the variables assigned in that hypothesis by H, and the set of hypotheses (treating each set of assignments as a separate hypothesis) as $\mathcal{H}$.

The first question a formal account of explanation must answer is which variables should be used in constructing $\mathcal{H}$. One possibility is for every explanation to include an assignment for every unobserved variable. However, Bayesian networks often use variables not meant to correspond to real entities in the world (e.g., a noisy-or gate for combining the influence of two causes). Additionally, there are often many variables that are not invoked in an explanation, and so a notion of "relevance" can be useful, allowing assignments to a subset of the unobserved variables (or even variables that are observed but not in O_{exp}).

Some models first generate $\mathcal{H}$ and then evaluate each hypothesis and rank them accordingly. Others "grow" their hypotheses by iteratively adding variables based on their ability to improve the explanation, stopping when the hypothesis cannot be improved further[18,19]. The hypotheses under consideration can then be evaluated and ranked, but note that what counts

as an improved hypothesis and what counts as a better explanation can be based on different criteria even within the same model. Some models aim to maximize the probability of the hypothesis given the observations ($p(h|\hat{o})$)[15,16]. Some models are more concerned with other metrics, such as the relative likelihood of the observations under one hypothesis ($p(\hat{o}|h)$) compared to the rest of the hypothesis set[42,21]. And some models aim to maximize how much information is gained about the explanandum were the hypothesis assumed or made to be true[18,19].

We now introduce the four models that we consider in this chapter — Most Probable Explanation[15], Most Relevant Explanation[42], Explanation Trees[18], and Causal Explanation Trees[19].

1.2.2 Most Probable Explanation (MPE)

Most Probable Explanation (MPE) ranks highly hypotheses with the most probable assignments to all unobserved variables, conditioning on $\hat{O}$. That is, every h in $\mathcal{H}$ includes an assignment for every variable in $\mathcal{X} \setminus \hat{O}$.[‡] This model leverages the intuition that the best explanation is one that is most probable given what we have observed[15]. The result is

$$MPE = \arg\max_{h \in \mathcal{H}} p(h|\hat{o}). \tag{1.1}$$

1.2.3 Most Relevant Explanation (MRE)

Rather than choosing the hypothesis that maximizes the probability of the unobserved variables given the observed values, we could choose values for the unobserved variables to maximize the probability of the observations ($\arg\max_{h \in \mathcal{H}} p(\hat{o}|h)$). Methods that pursue this route are known as likelihood models.

One problem faced by likelihood models is that multiple hypotheses will sometimes give the same high probabilities to the observed data[19]. For example, consider the case where we know the structure of a causal system like the circuit in Figure 1.1 from Yuan and Lu[42]. Likelihood methods would treat any hypothesis containing a union of A, "B and C", or "B and D" as equally good — the current flows equally well (perfectly), regardless of the particular path it

[‡] We might allow $h \in \mathcal{H}$ to include only those variables that are relevant for explaining $\hat{O}$. This is known instead as the maximum a posteriori model. There are a variety of possible relevance criteria as explored by De Campos et al.[17], but this problem is substantially more computationally complex than MPE. Here, we focus on MPE.

takes. This can make it difficult to choose between these explanations within the likelihood framework.

Rather than maximizing the likelihood per se, we can instead choose the hypothesis, h, that has the highest likelihood *relative* to the summed likelihood of all the other hypotheses in $\mathcal{H}$ except for h:

$$\frac{p(\mathcal{O}|h)}{\sum_{h_j \neq h, h_j \in \mathcal{H}} p(\mathcal{O}|h_j)}. \tag{1.2}$$

Yuan and colleagues' Most Relevant Explanation (MRE) model[42,21] proposes that the best explanation maximizes this quantity. This term plays an important role in statistics, known as the Generalized Bayes Factor[44], and in psychology, as a measure of how *representative* some data is of a hypothesis[45,46].

1.2.4 Tree-based models: ET and CET

The methods we have explored so far presume that you have $\mathcal{H}$ and then evaluate each hypothesis to determine which is best. However, in cases where the variable set is large, this can be difficult and computationally prohibitive. A class of *tree*-based models addresses this problem by using an iterative process for arriving at explanations. These models construct an explanation piece-wise, adding variables to the hypothesis one at a time, by choosing the best variable, assigning the variable a value and repeating until no further gains can be made. The resulting hypotheses are then evaluated based on some criteria, producing a list of explanations ranked by their goodness. Models differ in how they choose the best variable to add, how they decide to stop, and how they then evaluate the resulting hypotheses.

1.2.4.1 Explanation Trees

The Explanation Tree (ET) model — as proposed by Flores et al. [18] — determines which variable carries the most information about the rest of the unknown nodes, conditioned on what is already known. In ET what is already known includes $\hat{O}$ and any variables included in hypotheses farther up the tree. This means that at the beginning (when the hypothesis is $\varnothing$) the model selects the node that provides the most information about the rest of the unobserved variables conditioned on $\hat{O}$. Formally, we grow h' (the hypothesis up to that point) by choosing the X_i as the maximum of $\sum_Y \text{INF}(X_i; Y|\hat{O}, h')$, where Y is shorthand for $\mathcal{X} \setminus \{\hat{O} \cup h' \cup \{X_i\}$, or all of the variables not observed, included in the current hypothesis or currently under con-

sideration, and INF(·) is a metric of informativeness. In our calculations we will use *mutual information* as our INF(·), as in Nielsen et al. [19].[§]

Once a variable is chosen, each assignment creates a new branch, and that assignment is added to the interim hypothesis h', and is effectively treated as an observed variable. The process is then repeated until adding any more variables is deemed to provide a hypothesis with a probability that is too low, as defined by parameter β_{ET}, or to carry too little information, as defined by parameter α_{ET}. This process provides multiple, mutually exclusive explanations that can vary in their complexity based on how much information the complexity buys.[¶] Once these hypotheses are assembled, the model ranks the explanations by the posterior probability of each branch of the tree – i.e., how likely each hypothesis is, given the observed data.

Up to this point every model we have considered assumes the set of observed data is the data we are explaining, or $\widehat{O} \equiv O_{\text{exp}}$. The ET model further assumes that we aim to reduce uncertainty of the entire variable set $\mathcal{X}$ in deciding which variables are ostensibly relevant to our explanandum, O_{exp}. However, these assumptions can be problematic. For example, in ET, a variable that is unrelated to O_{exp} but carries a lot of information about other unknown variables may be added to the hypothesis despite its irrelevance to our explanans.

1.2.4.2 Causal Explanatory Trees

The Causal Explanatory Tree (CET) model introduced by Nielsen et al. [19] addresses these weaknesses. Rather than using traditional measures of information such as mutual information, CET uses *causal information flow* [29] to decide how the tree will grow. Causal information flow uses the post-intervention distribution on nodes (as proposed in Pearl [28]) rather than considering the joint probability distribution "as is". To extend Ay and Polani's [29] analogy, imagine pouring red dye into a flowing river. You could identify which way is downstream by tracking the red streak that results; if you were to pour in the dye just after a fork in the river, you would not find red dye in the other half of the fork. Now consider the case of a static, dammed river — a river that does not flow. If you poured the dye just after the fork, redness would gradually diffuse through the water, eventually reaching the other path from the fork and tinting the whole river. In this case, there is no concept of something being 'downstream'. Causal information attempts to capture the notion of 'downstream' influence that is absent in traditional mutual information.

[§] Flores et al. [18] consider several versions of INF.

[¶] Mutual exclusivity refers to the fact that once a variable is assigned, it holds through the rest of the tree.

We denote post-intervention distributions with a "$\hookrightarrow$" on a conditioned variable [⋆]. If we have variables W, X, Y, Z, where we have observed $W = w$, intervened on Z (giving us post-intervention values $\bar{Z} = \bar{z}$), then the causal information passed from X to Y is,

$$\sum_{x \in X} p(X = x | W = w, \bar{Z} = \bar{z}) \times \sum_{y \in Y} p(Y = y | \bar{X} = \bar{x}, w, \bar{z}) \log \frac{p(y|\bar{x}, w, \bar{z})}{p(y|w, \bar{z})}. \quad (1.3)$$

This allows us to specifically ask the degree to which a variable ($X \equiv X_i$) influences the explained data ($Y \equiv O_{\text{exp}}$), treating the non-explained data as observed ($W \equiv O_{\text{not-exp}}$) and previous parts of the explanation as intervened on ($Z \equiv h'$). This solves the problem of distinguishing between explained and unexplained observations ($W \neq Y$). It also allows us to maximize information about the O_{exp} rather than $\mathcal{X} \setminus \hat{O}$ as in ET. However, like ET, the CET model proposes variables iteratively, until no remaining variables add more causal information than the criterion α_{CET}. Then each branch is assigned the score $\log\left(\frac{p(O_{\text{exp}}|\bar{h}', O_{\text{not-exp}})}{p(O_{\text{exp}}|O_{\text{not-exp}})}\right)$ where $\bar{h}'$ is the total set of assigned values in a hypothesis at a branching point.

1.3 Comparing model and human explanation judgements

We now compare the prediction of these four models against human judgements when both generating and evaluating explanations. We focus on explanations in the two Bayesian networks shown in Figure 1.1. The *Pearl* structure is derived and parameterized as in Nielsen et al. [19]; the *Circuit* graph and its parameters are taken from Yuan and Lu [42]. These networks have been used previously to distinguish between the performance of different models. Each network consists of several binary variables, prior probabilities on those variables, and relationships between variables. We consider the case where only one variable is observed, in *Pearl* $D = 1$ and in *Circuit* $O = 1$, and these act as both $\hat{O}$ and O_{exp}, i.e., each is the only variable we observe and explain in that structure.

The models diverge in how they rank explanations in *Pearl* and *Circuit*. In past research, the *Pearl* structure was used by Nielsen et al. [19] to argue in favour of the CET, and the *Circuit* structure was used by Yuan and Lu [42] to argue in favour of the MRE.[✠] By drawing from distinct research lines we aim to be as fair as possible in testing the models.

[⋆] In other chapters, we will use ☞ to indicate an intervention.

[✠] The CET had not been published by the writing of Yuan and Lu [42]. Yuan et al. [21] addresses CET but that work involves more complicated scenarios than those considered here.

In addition to being useful for distinguishing between models, these structures have properties that are particularly interesting from a psychological perspective. The *Pearl* structure includes complex causal dependencies that cannot be easily captured by the paradigms used in cognitive psychology. The *Circuit* structure contains explanations with equal (perfect) likelihoods for the observation, but which vary in the number of variables cited in the explanation. Research on people's preferences for simplicity bear on this case, which shows that people may choose an explanation with fewer causes even if it is less likely than other more complex alternatives[35].

In the past, researchers used the match between their own explanatory intuitions and the models' predictions to provide support for their model. However, this method can be problematic: Nielsen et al.[19] and Yuan and Lu[42] conflict in their intuitions, leaving us in a quandary. We generalize the intuition-matching approach using two experiments in which we ask people to generate (Experiment 1) and evaluate (Experiment 2) explanations in cases formally equivalent to *Circuit* and *Pearl*. We used MPE, MRE, ET, and CET to rank the quality of various explanations, and analyse these rankings as they compare to the rankings derived from human explanations. By appealing to a wider array of human judgements we hope to extricate ourselves from this quandary.

1.4 Experiment 1: Generation

1.4.1 Participants

We recruited 188 participants through Amazon Mechanical Turk; 9.6% of those failed to complete the study, did not consent to taking the study, or did not follow the instructions, and 35.9% failed at least one explicit reading/attention check. This left 109 participants for analysis ($M(\text{age}) = 27.7$, %-Female $= 29.3\%$).

1.4.2 Materials & procedure

Participants were randomly assigned to either the *Pearl* or *Circuit* structure. They then were assigned to one of two semantically-enriched stories embodying a causal structure, involving either novel alien diseases or the ecology of lakes. For example, one of the two scenarios adapted from the *Circuit* structure taught participants about the effects of novel diseases on producing a kind of fever.

For this scenario, participants received facts about the base rates of four novel diseases (corresponding to $p(A)$, $p(B)$, $p(C)$, and $p(D)$), and information allowing them to understand which diseases would produce the fever, which would only occur in the presence of two proteins X and Y. One disease (corresponding to A) produced both the necessary proteins and thereby caused the fever. The second disease (corresponding to B) produced one of these proteins, and when paired with either the third and/or the fourth diseases (i.e., C or D) which produced the other protein, would be sufficient to cause the fever. X and Y were added to provide an intuitive mechanism outside of the domain of circuits that describes the complexities of *Circuit*'s causal relations. Probabilities were presented as frequencies (out of 1000) and act as realizations of the probabilities in the graphs in Figure 1.1.

In order to ensure that participants were paying attention, we asked questions that required simply reading the information off a figure (e.g., "Out of 1000, how many aliens have [disease A]?"). Participants who failed any comprehension questions were excluded from subsequent analyses. To ensure that participants' judgements were not limited by memory, the base rates and causal structure were available when answering these reading checks as well as during the generation portion of the experiment. Participants were asked to use the information that had been provided to write down "the **SINGLE BEST EXPLANATION**" for the observed effect (e.g., for a particular alien's fever), where "a 'single' explanation can include more than one causal factor." Participants were explicitly asked not to list multiple possible explanations, but rather to "identify the one explanation that you think is the best." This was meant to exclude what we call "disjunctive" explanations like "It was A or B and C and not D", or, formally, as $A = 1 \cup \{B = 1 \cap C = 1 \cap D = 0\}$.

1.4.3 Results and discussion

Participants' explanations were coded by an assistant blind to the authors' hypotheses. The coder's goal was to identify which variables were mentioned and what values were assigned to those variables. We excluded participants who gave a response that conflicted with our instructions, such as providing a disjunctive explanation.

In *Circuit*, most participants provided explanations that fell into one of two options: BC (43%) or A (40%), and, in *Pearl*, most participants chose one option: they attributed the disease to the presence of a genetic risk factor and not receiving the vaccine (73%, see Figure 1.1).

For the explanations participants generated, we computed measures of explanation quality under each of the four models and saw which models gave better scores to those explanations

Table 1.1: Rank-correlations for models and human data in Experiment 1, $p < 0.05$ in **bold**, < 0.10 in *italics*.

	Circuit		*Pearl*	
Models	ρ_{Spearman}	P_{val}	ρ_{Spearman}	P_{val}
MPE	-0.06	0.631	0.32	0.449
MRE	*0.20*	*0.074*	**0.83**	**0.017**
ET	0.08	0.460	0.17	0.700
ET$_{\text{tree}}$	0.01	0.900	0.41	0.310
CET	*0.22*	*0.055*	**0.93**	**0.003**
CET$_{\text{tree}}$	0.06	0.590	**0.77**	**0.032**

that were generated more frequently. This process provides us a rank for each participant's explanation according to each of the models and a rank of how frequently each explanation was generated, which allows us to calculate a Spearman rank-order correlation between participant's aggregate explanation choices and the models' predictions, see Table 1.1.

Note, we used two versions of the *tree*-algorithms: one where explanations not reached by the tree received the lowest possible rank (which we give the subscript "$_{\text{tree}}$"), and one where we ignored these exclusions and applied the evaluation criteria used at each branch point. The tree models were designed to both generate and evaluate explanations "on the fly", but it is not clear whether the way models *generate* explanations has led to their success in previous literature. Model success (or failure) may be the result of the branch evaluation criterion, rather than the result of the algorithm for generating hypotheses. This is why we analyse these parts of the algorithms separately.

We find that MRE and CET are most consistent with participants' judgements (though they still only reach marginal significance in the *Circuit* case). In contrast, for both structures, models that rely only on an assignment's probability (i.e., MPE and ET) poorly predict the explanations that people generate (in *Circuit*, MPE had a negative coefficient).

The major weakness of the tree versions of CET and ET lies in the fact that once a node is chosen for expansion, it remains expanded. Thus, mutually exclusive explanations cannot be reached in the same tree. That is, in *Circuit*, A and BC were the two most popular explanations and $A \cap BC = \varnothing$, so the first step to include either A or B will preclude the other explanation. Empirically, participants are roughly split between these two explanations, which suggests that any method that generates a unique best explanation will always fail to capture the

variability that results when people are generating explanations, even if those people are generating explanations about the same system. We studied only deterministic algorithms which may be causing the models to diverge from people in how they generate hypotheses. Adding probabilistic rules may also be important for accounting for uncertainty about the parameter estimates, which in the real world are typically not given to you but must be inferred from data as well.

Note that CET in this case treats all explanations that sufficiently determine the observations as having equivalent rank. Because the system is deterministic, all 38 of the sufficient explanations are ranked as number 1 — or rather, because they are so numerous, number 19. This is a problem unique to CET, and results from its use of intervention, which ignores variables' prior distributions in determining an explanation's score.

1.5 Experiment 2: Evaluation

In Experiment 1, we found evidence that at least some of the proposed models capture people's explanatory intuitions. Of course we should have expected some of the models to perform well; what is remarkable is how poorly some of the models did. In particular, we saw surprisingly poor performance from the tree-growth models as compared to their exhaustive-search evaluative counterparts.

Generating explanation is harder than only evaluating them — generation requires searching through the hypothesis set and then evaluating the generated explanations, while evaluation only requires computing a known evaluation function. The tree versions of the tree models are designed to make generation tractable. However, if complexity were the primary hurdle, in *Circuit* where the hypothesis space was much larger, we would expect tree methods to perform comparatively better than in *Pearl*. But they were relatively *worse*. This was due to the fact that the tree models were guaranteed to cut off at least 40% of participants since A and BC were the top choices, and cannot be reached in the same tree.

It is striking that methods that relied on probability (MPE and ET) performed so poorly in contrast to MRE and CET. However, these results may only apply to situations in which explanations are generated; explanations with large absolute probabilities may be difficult to access when generating explanations but could still be preferred if people only need to evaluate predefined hypotheses. There are many cases in which a hypothesis proves incredibly hard to generate, but once generated quickly becomes welcomed as the best explanation for many phenomena (e.g., Newton's and Einstein's physics). And, if conquering search problem is one of

the driving factors behind the success of MRE and CET, it is possible that they could fail in the evaluation case.

In order to test these ideas, we conduct an experiment that is almost identical to Experiment 1. But, rather than asking people to generate explanations, we take that burden off of their shoulders. Instead, we ask them to evaluate a set of explanations that we generate for them.

1.5.1 Participants

A total of 245 participants were recruited through Amazon Mechanical Turk, with 9.8% excluded for failing to provide consent or otherwise complete the study and 25.3% excluded for failing one or more reading checks. This left 165 participants for analysis ($M(\text{age}) = 31.3$, %-Female $= 34\%$): 46 in the disease version of *Circuit*, 46 in the lake version of *Circuit*, 34 in the disease version of *Pearl*, and 39 in the lake version of *Pearl*.

1.5.2 Stimuli

An explanation was included in the study if either criterion held:

- The explanation was generated by more than one participant in any one condition in Experiment 1.
- The explanation was in the top two explanations generated by any of the models.[⊎]

This yielded thirteen explanations for the *Circuit* causal structure and six for the *Pearl* causal structure.

1.5.3 Procedure

The materials and methods were nearly identical to those in Experiment 1, with the following important change: instead of providing an explanation, participants were asked to rate the quality of several provided explanations. Specifically, they were asked to rate each explanation

[⊎] Because there are many ways one can interpret what counts as one of the two "top" explanations, we allowed the top two as defined by *any* interpretation found in the literature of how to rank a model's results. For example, Yuan [20] and Yuan et al. [21] include only minimal explanations (i.e., explanations for which no subset has appeared prior to it in the ranking of explanations) when determining the results of MRE, whereas Nielsen et al. [19] simply listed explanations based on their scores regardless of their minimal or non-minimal status.

"by placing the slider next to each explanation along the spectrum from Very Bad Explanation (furthest to the left) to Very Good Explanation (furthest to the right)," where intermediate ratings could fall anywhere in between.

Although the sliders were not presented with a numbering, positions implicitly corresponded to values between 0 and 100. Based on these ratings we can again create an explanation ranking for each participant, with ties being treated as in Experiment 1 as a repeated average value. By using ranks rather than continuous ratings we need only assume that participants have a monotonic relationship between bad and good, and avoid making assumptions about the particular nature of that scale for each participant.

1.5.4 Assessing model predictions

For each model, we calculated the scores assigned to the explanations that were provided to human participants. Because we were interested in explanation evaluation, we did not limit the ranks derived from CET or ET to those generated by the trees, but we did limit MPE to complete assignments, as otherwise it would be equivalent to ET.

To generate scores indicating the quality of each model, we created a set of intersection proportions. To illustrate, were we to consider only a single participant, this involves the following process. We take the human ranking as the veridical ranking. We then check whether the model's top rank explanation is the same as the participant's. We then check whether the model's two highest-ranked explanations are included in either of the two highest-ranked human explanations. We continue to do this for the whole explanation set, identifying the number of model explanations that were ranked at a level less than or equal to each level of human ranking. We can repeat this with every participant, to obtain the number of explanations matched at each rank for each participant. We can then take the average of these scores at each rank, giving us the intersection size for the full population.

It is important to note that the absolute intersection size is less useful than the proportion when we are comparing between causal structures. We can transform these values into intersection proportions by dividing each value by the total number of model explanations. This maps to a measure of how many of the model's top explanations are thought by the models to be at least as good as those generated by the average person up to that point.

To illustrate, suppose that we had explanation set $\mathcal{H}$: A, BC, BD, $ABCD$, and $\not\!B$, and we were considering a participant(P) with a ranking of $P(1) = BC$, $P(2) = A$, $P(3) = ABCD$, $P(4) = BD$, and $P(5) = \not\!B$. To compute a model's performance, we

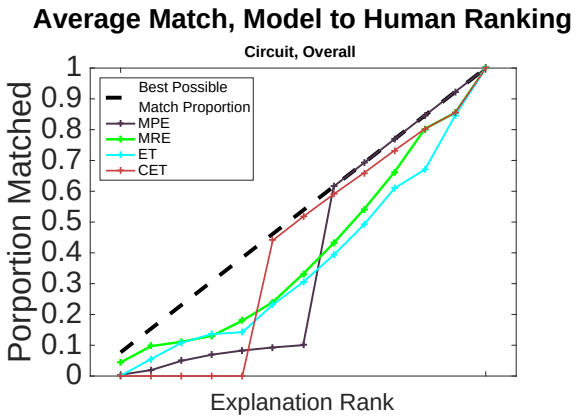


Figure 1.2: Results for Experiment 2: Average intersection proportions for *Circuit* conditions.

would look at the ranking that the model(M) assigned to the different explanations. If their top ranks matched, i.e., $M(1) = BC$ was the model's top choice, then the first value would be $V(M, P, 1) = \frac{1}{5} = \frac{|M(1)=\{BC\} \cap P(1)=\{BC\}|}{|\mathcal{H}|}$, and if it was not the score would be 0 since $M(1) \cap P(1) = \varnothing$. This process would be repeated for the first and second values, i.e., the next value is $V(M, P, 2) = \frac{|\{M(1)M(2)\} \cap \{P(1)P(2)\}|}{|\mathcal{H}|}$, and so on until we got to $V(M, P, 5)$ which will necessarily equal 1 since both rankings were defined relative to the same set, meaning the two sets are equivalent and are also both equivalent to $\mathcal{H}$.

Figure 1.2 displays the intersection proportion for the *Circuit* structure, and Figure 1.3 displays those for the *Pearl* structure.

Another method for capturing overall model performance is to take the sum of the average values at each point. The best one can do in the intersection proportion is to match every explanation up to that rank at each rank. A perfect summary score is, $\sum_{i=1}^{|\mathcal{H}|} i/|\mathcal{H}|$. For *Circuit* the maximum summed intersection value is $\sum_{i=1}^{13} i/13 = 7$ and for *Pearl* it is $\sum_{i=1}^{6} i/6 = 3.5$.[⋉] These values can be found in Table 1.2.

1.5.5 Results and discussion

As you can see in Figures 1.2 and 1.3, both MRE and CET are closer to the dotted line in general, i.e., they are better on average than either MPE or ET.

One interesting pattern to note is a trend that echoes results for CET in Experiment 1. CET

[⋉] One could think of this as an estimate of the area under the curve defined by the intersection proportions.

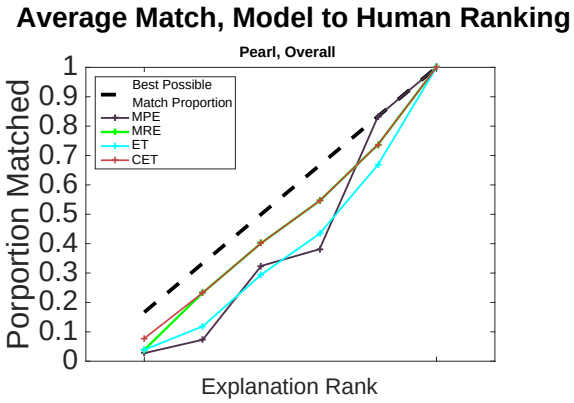


Figure 1.3: Results for Experiment 2: Average intersection proportions for *Pearl* conditions.

Table 1.2: Summed intersection values for models.

Models	*Circuit* Score	*Pearl* Score
MPE	5.26	2.64
MRE	5.43	2.96
ET	4.99	2.55
CET	**5.60**	**3.00**
Max Value:	7	3.5

stays flat at zero for a while and then rapidly accelerates as it goes forward. This is a consequence of the interaction between CET's reliance on intervention and the deterministic causal system in the *Circuit* condition. Because so many of the explanations are sufficient for bringing about the effect in question, many explanations share the role of the 'best' explanation. And because we choose an explanation's rank in the case of a tie as the average rank of all those in the tie had they not been in a tie, many of the best explanations are given a fairly high value. Thus, once we get to the sixth item, $M(1)$–$M(5)$ have had equal scores to $M(6)$, and once the values pass that threshold CET's $V(M, P, \cdot)$ rapidly catches up to and passes MRE's (which was otherwise in the lead). MPE, on the other hand, has the opposite problem: only two of its values are defined and so the other eleven explanations all receive a score of 8, resulting in perfect performance from 8 onwards (though most of its ranks are, by definition, undefined).

Table 1.2 shows that in both structures CET does the best, followed by MRE, then MPE and

finally ET.

1.6 General discussion

We began this chapter with the aim of systematically evaluating formal models of explanation against human intuitions as well as clarifying human explanation through the lens of computational models. We consider how our results address these aims.

1.6.1 Evaluating models of explanation

We find that CET and MRE provide reasonable but imperfect fits to human judgements in both the *Circuit* and *Pearl* structures, and for both explanation generation and evaluation. MPE and ET perform less well. This suggests that human explanation is not explained well by appealing to maximum posterior probability values. Instead, it seems that a measure of evidence (MRE) or causal information (CET) may better model human explanation.

These findings indicate that the algorithms used for generating explanations in the tree methods (ET and CET) fail to capture an important aspect of human intuitions about explanation — explanations that are radically different from one another (i.e., that cannot be reached by the same tree) may both be seen as valid explanations. In the generation task, the purely evaluative tree models outperformed their generative counterparts. The evaluation function seems to be quite important, but it has been emphasized less than the generation algorithm in previous work[18,19]. The evaluation function merits closer inspection.

Speaking generally, our work reveals the degree to which a model's objective alters that model's predictions. Our analyses highlight the problem with using hard intervention in deterministic cases. CET gave the same score to all 38 sufficient explanations that, presumably, we would want the model to distinguish. MPE and ET excel at doing what they were created to do, but we may wish to distinguish between their goals (which do not correspond closely to human explanation judgements) and the goals of models like CET and MRE (which do).

1.6.2 Bidirectional implications from human and formal explanation

These results indicate that formally characterizing the objective function implicit in human explanation may be a challenging but exceptionally useful task. The variability in how well these formal models performed demonstrates that despite seeming straightforward, how people choose a good explanation has many hidden subtleties and complexities. The good perfor-

mance of CET and MRE relative to MPE and ET suggest that human explanation is likely more concerned with causal intervention or the relative quality of a hypothesis than it is with absolute judgements of posterior probability. But the alternative hypothesis set and the role of intervention have received relatively little attention in psychological research on explanation. On the other hand, simplicity was not explicitly represented in the formal models we explored (but, see De Campos et al. [17]), but has been found to affect human explanatory judgements [43]. Then, it is surprising that a large proportion of people explain using BC over A in the *Circuit* example, when BC is both less likely and more complex than A. Probability, simplicity, intervention and alternative hypotheses seem to weave a rather complex image — an image just asking to be unravelled.

All of the models we studied require knowing a priori the causal structure and parameterisation, whereas people must infer these values from finite amounts of data. Though explanation has been tied to improved learning, we know much less about how the learning process and the processes for generating and evaluating explanations interact with one another. Additionally, developing extensions of these models that can learn from finite amounts of data will increase the expressiveness of the models while also making them more able to deal with the problems that both humans and many real intelligent systems face.

1.6.3 Conclusion

Given that explanation plays an important role in human inductive judgements [43], where humans still outperform artificial systems, we propose that models will benefit from a closer match to human judgements. And conversely, given that formal models need to make explicit the roles played by different parts of the explanatory problem and its solution, we propose that psychological accounts of explanation will benefit from models that precisely specify formal characteristics for what makes a good explanation. Both inquiries benefit from attending to the other. Our work, in simultaneously analysing models of explanation from artificial intelligence and the psychology of human explanation, embodies this view.

2

Ockham's Razor Cuts to the Root: Simplicity in causal explanation*

2.1 Introduction

Simpler explanations are better explanations. This intuition, right or wrong, often guides both scientific and everyday reasoning, earning the moniker "Ockham's Razor" for the unnecessary complexities that it "cuts" out of explanations. While simplicity is lauded by both scientists and philosophers, there is little consensus on how simplicity should be defined. William of Ockham argued that we "not multiply entities beyond necessity," suggesting that simplicity is a matter of the *number of entities* involved in an explanation. Newton's first Rule of Reasoning in Philosophy is "we admit no more causes …than [those] true and sufficient to explain [our observations]", suggesting *causes* are the units in which simplicity is measured. Einstein tells us that "the grand aim of all science … is to cover the greatest possible number of empirical facts … from the smallest possible number of hypotheses or axioms" (for quotations, see Baker [26]), suggesting the size of a *set of hypotheses* or *axioms* is what matters.

Beyond these classic examples, contemporary philosophers, statisticians, and computer sci-

* Some of the content in this chapter is derived from a paper currently under review that was co-authored with Tania Lombrozo.

entists have developed formal definitions of simplicity that can be used to guide theory choice and inference (for review see Sober[47]). Simplicity is argued to lead to more accurate inference[24], better predictions[48], or more efficient learning[49]. These proposals are often grounded in algorithmic information theory and probability[22], Kolmogorov complexity[23], the cardinality of parameterized models[50,51], or the implicit size of the hypothesis space ("the size principle", see Tenenbaum and Griffiths[24]). Within psychology, these approaches to simplicity have proven useful in modelling perceptual classification[52], language[53], and the perception of hierarchically structured domains in general[27]. But, these different metrics vary in how well they fit real-world applications: for example, Kolmogorov Complexity can be easily applied when the problem can be described as predicting the next element in a sequence composed of characters from a fixed alphabet (e.g., predicting the next letter in the sequence "banan_"). But some scenarios cannot be easily framed as sequences of this type. Nonetheless, people reason about such scenarios.

Thus, though formal approaches to defining simplicity in well-specified domains have been fruitful, research on intuitive judgements of simplicity in everyday explanations has made considerably less progress. This is unfortunate, as explanation is a ubiquitous phenomenon. People constantly explain the social and physical world around them, and their explanatory choices have important consequences in a variety of domains[54,35,43]. For instance, explanations for our own and others' behaviour can affect judgements of responsibility and blame[55,56], and clinicians' explanations for a patients' behaviour can affect diagnoses and treatment decisions[57,58,59]). If people prefer simpler explanations in these domains – and there's reason to think that they do[60,61,62] – it's especially important to provide a more precise characterization of simplicity in explanations, and to better understand the implications of a preference for simpler explanations.

In this chapter, we consider the nature and role of simplicity in human judgement, focusing on the explicit evaluation of causal explanations, such as explanations for symptoms that appeal to underlying diseases. In four experiments, we address the following questions about simplicity in the context of causal explanation and its role in human cognition:

Q_1: What makes a causal explanation simple?

Q_2: How are explanations selected when the simplest explanation is not the one best supported by the data?

Q_3: What are the cognitive consequences of a preference for simpler explanations? For example, does the preference bias memory or inference?

Q_4: Why do people prefer simpler explanations?

We begin by differentiating two metrics for simplicity, node simplicity versus root simplicity, and motivate these questions in light of prior research. We then report four novel experiments.

2.1.1 Defining Simplicity: *Node* versus *Root* Simplicity

Ockham's razor canonically applies to arguments about the number of entities or the number of *kinds* of entities postulated to exist, a notion that is often referred to as "parsimony." In contrast, previous work on simplicity in causal explanatory judgements has typically focused on "elegance"[26], where the *kinds* of causes are known (e.g., which diseases exist), and competing explanations differ in which of these causes they invoke in a given case (e.g., stating that a disease is present to explain a given patient's symptoms). In this work, simplicity has been measured in terms of the number of causes invoked in the explanation[63,64,43,62,65] †. We call this metric *node simplicity*, as it involves counting the total number of causal nodes cited as being present in the explanation.

To illustrate node simplicity, consider the case of Chris, who has been *extremely fatigued* and has been *losing weight*. What explains these symptoms? Chris could have chronic fatigue syndrome, an explanation which invokes one cause to account for both symptoms. Another possibility is insomnia (to explain the fatigue) and a decrease in appetite (to explain the weight loss), thereby invoking *two* causes. On the grounds of node simplicity, the first explanation is preferable to the second — one disease is fewer than two diseases. Read and Marcus-Newhall[62] and Lombrozo[35] found that when the probabilities of the corresponding explanations were unspecified, participants preferred explanations consistent with this metric — that is, they preferred to explain multiple symptoms with the smallest number of diseases. However, both Lagnado[64] and Lombrozo[35] found that this preference was eliminated or tempered when the simplest explanation was not the most likely. In the case of Chris, chronic fatigue syndrome could in fact be less common than having the conjunction of insomnia and a decreased appetite (if, e.g., Chris happens to belong to a population of particularly sleepless and sated people).

Both Lagnado[64] and Lombrozo[35] investigated people's explanatory preferences in cases where simplicity and probability were in conflict, using disease examples similar to those de-

† Strictly speaking, Read and Marcus-Newhall[62] and Thagard[65] quantified simplicity in terms of the number of *propositions* involved in an explanation. However, in the stimuli used in Read and Marcus-Newhall[62], each proposition corresponded to the presence of a cause.

scribed above. Both researchers found that when a complex explanation was explicitly identified as more likely than a simpler alternative, participants chose the more probable explanation. However, Lombrozo [35] additionally examined cases in which participants were provided with more indirect probabilistic cues: the base-rate of each disease. While this information was sufficient to evaluate the relative probabilities of the explanations (under assumptions about independence between the diseases), participants' choices were nonetheless influenced by simplicity. In particular, participants had an overall preference for the simpler (one-cause) explanations, but this preference was tempered by probability information. This generated a pattern of judgements consistent with the interpretation that simplicity altered the prior probability assigned to explanations, with very strong probabilistic evidence required to overcome this initial bias. Bonawitz and Lombrozo [63] found a similar pattern of results in preschool-aged children.

This previous work establishes that simplicity is a powerful force in determining explanatory preferences, but no empirical research (to our knowledge) has attempted to differentiate alternative metrics for simplicity in explanation choice. This is problematic given that prior results are not uniquely consistent with node simplicity. We propose an alternative metric that can also explain these results, which we call *root simplicity*. Informally, root simplicity can be defined in terms of the number of *assumed* or *unexplained* causes in an explanation, where simpler explanations are those with fewer assumed or unexplained causes (which, for present purposes, we treat as interchangeable).

This metric is related to a number of proposals from philosophy and the history of science concerning the value of simplicity and the goals of scientific theorizing, though they have not always been expressed in terms of root causes. For example, the quote from Einstein included above indicates a preference for a small number of axioms, where axioms are similarly "assumed" or unexplained. Relatedly, Friedman [66] endorses explanations that *unify* phenomena with few assumptions, saying "science increases our understanding of the world by *reducing* the total *number of independent phenomena* that we have to accept as ultimate or *given*" (emphasis added). While Friedman has in mind explanations for different types of properties or events, the number of independent phenomena that are given or assumed maps roughly onto the number of causes that are given or assumed in explaining a token event, which corresponds to root simplicity.

Although the materials from Read and Marcus-Newhall [62] and Lombrozo [35] do not differentiate between *node* and *root* simplicity (both metrics predict the same judgements), there are other cases for which these two metrics diverge. To illustrate, consider that depression is a known cause of *both* insomnia and loss of appetite, and suppose that we know that Billy does

not have Chronic Fatigue Syndrome. This leaves us with the following two explanations for Billy's fatigue and his decreased appetite: insomnia and loss of appetite, which were themselves caused by depression, or insomnia and loss of appetite, which were not caused by depression and instead arose independently (see Figure 2.1). We call the first explanation the complete-choice because it includes the complete set of possible causes, and the latter explanation the proximal-choice because it includes only the most proximal causes (i.e., only the causes that directly generated the tiredness and weight-loss).‡

In this scenario, node and root simplicity diverge. Node simplicity would say that the complete-choice has a measure of three (because it cites all three causes) and the proximal-choice a measure of two (because it cites two causes). Thus, if people employ node simplicity in evaluating explanations, they should prefer the proximal-choice explanation. However, according to root simplicity, the complete-choice has a measure of one (because we only assume that Billy is depressed) while the proximal-choice has a measure of two (because it assumes that Billy independently developed both insomnia and a reduced appetite). Root simplicity, in contrast to node simplicity, favours the complete-choice explanation.

As a second example of a scenario for which node and root simplicity generate divergent predictions, consider two candidate explanations for a heart attack. In one case, the cause is heart disease, which is itself caused by metabolic syndrome (the complete-choice explanation). In the second case, the proximal cause is heart disease, but where the heart disease was not caused by metabolic syndrome — it is itself assumed or unexplained (the proximal-choice explanation). In this case, node simplicity favours the proximal-choice explanation (one cause is fewer than two), while root simplicity does not predict a preference for either explanation (in both cases, the causal chain has one assumed cause).

Examples like these, for which the two metrics predict different preferences, allow us to investigate whether node or root simplicity better characterizes people's explanatory preferences and thus address our first research question ($\mathbf{Q_1}$): what makes an explanation simple? Moreover, by varying the probabilistic evidence for different explanations using structures like those

‡ We are contrasting the cases where something is present and where something is *not* present. Thus, ours is a discussion about a *sharp* Ockham's razor, which actively states that variables are not present, as opposed to a *dull* Ockham's razor, which is silent as to the presence or absence of variables[47]. This assumption plays a role in our later analyses, which involve comparing evidential support for different explanations. A hypothesis consistent with a *dull* Ockham's razor will include the possibility that the variables in question are present, and (assuming it is possible that the variable is not present) will always have greater probability than the hypothesis that the variable is present. This distinction echoes the debate between Popper[67] and Jeffreys[68] on simplicity in the context of the prior probability of various statistical models (c.f., Baker[26]).

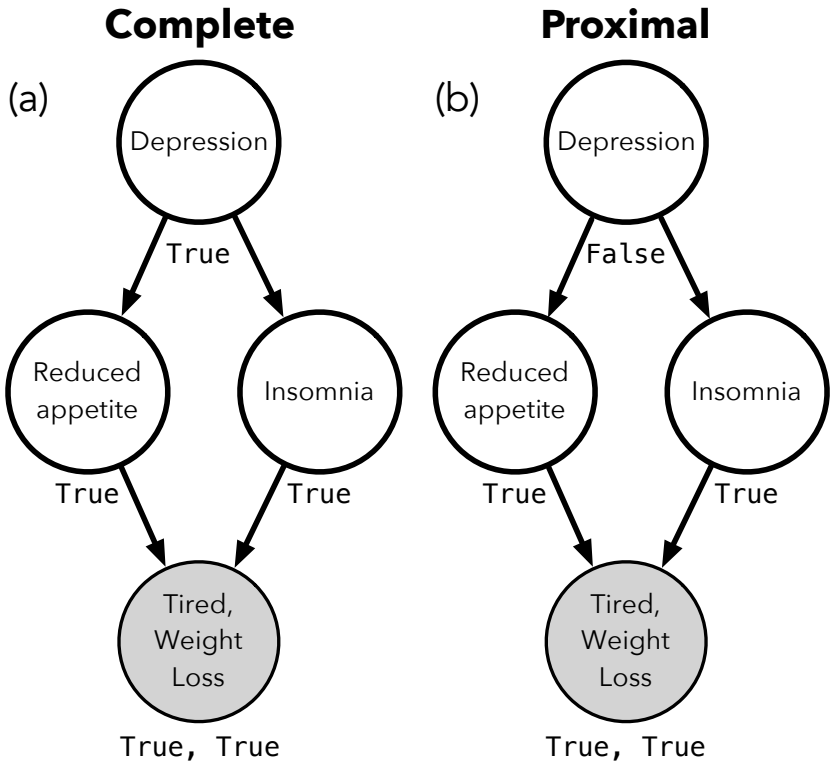


Figure 2.1: Illustrations of graphs corresponding to the Complete (a) and Proximal (b) explanations. Each circle is a variable or set of variables (e.g., disease or symptom set). The value of the node is indicated below the node; in this case nodes have values of present or not present. Arrows indicate potential causal relationships. Grey-filled circles indicate that the node's value has been observed.

just discussed, we can address our second question ($\mathbf{Q}_2$): how are explanations selected when the simplest explanation is not the one best supported by the data? These questions are the focus of Experiments 1–2.

2.1.2 Cognitive Consequences of a Preference for Simpler Explanations

What are the implications of a preference for simpler explanations? Previous research has shown that the act of explaining can impact both learning and inference (e.g., Koehler[69], Sherman et al.[70]; for review, see Lombrozo[43]). Lombrozo[35] found that participants who preferred an unlikely but simple explanation overestimated the observed frequency of the disease invoked in that simple explanation. However, Lombrozo[35] did not differentiate node and root

simplicity or go beyond an association to demonstrate a causal relationship between the act of explaining and the systematic estimation errors exhibited by some participants. Here, we vary the order in which participants explain and estimate to isolate the causal influence (if any) exerted by explanation on estimation. This allows us to address one aspect of our third research question ($\mathbf{Q_3}$): what are the cognitive consequences of a preference for simpler explanations? This is the focus of Experiment 3.

2.1.3 Why Do People Prefer Simpler Explanations?

Finally, why do people prefer simpler explanations? One possibility is that a preference for simpler explanations is just a human failing — perhaps a mostly harmless side effect of limited cognitive resources. Another possibility, however, is that favouring simpler explanations serves a useful cognitive function. This possibility is suggested by arguments in favour of simplicity in philosophy and statistical inference[26,68,67,71] — some even arguing that simplicity is a foundational principle through which all of cognition can be understood[72,25]. However, even among those who agree on simplicity's value, it serves no single, agreed-upon role. Different roles have been proposed, and each proposal constrains (and is constrained by) the metric used to define "simplicity" (e.g., Akaike[50], Chater[72], Jeffreys[68], Kelly[49], Popper[67]).

The possibility we explore is that one function of explanation is to facilitate the formation of relevant, information-rich representations of causal systems, where these representations are tailored to aiding future intervention and prediction in a variety of situations more general than the set of scenarios for which the explanation was originally invoked[73,74]. If this is the case, a preference for simpler explanations could exist to support the acquisition or deployment of these representations. We revisit these ideas in Experiment 4, where we tackle our final question ($\mathbf{Q_4}$): why do people prefer simpler explanations?

2.2 Experiment 1: Node versus Root simplicity

In Experiment 1 we test the predictions of node simplicity versus root simplicity against human judgements. Participants learn one of two causal structures involving novel diseases and are asked to provide the most satisfying explanation for an individual's symptoms. The causal structures are designed to support two alternative explanations for which node and root simplicity generate divergent rankings. In Experiment 1 we do not provide information about the

relative probabilities of different explanations. However, we introduce this information in Experiment 2.

2.2.1 Methods

2.2.1.1 Participants.

Sixty-eight participants were recruited online using Amazon Mechanical Turk and paid $.60 for their participation. Of these, 53% passed reading checks described below, leaving 36 participants for analysis. Participation was restricted to individuals with IP addresses from the USA and with HIT approval ratings of 95% or higher.

2.2.1.2 Materials and Procedure.

Participants were asked to imagine that they were doctors on an alien planet, Zorg. Their task was to assist in the diagnosis of alien diseases. Participants read information about the causal relationships between diseases that afflict the aliens living on Zorg. This causal information varied across the Diamond-Structure and Chain-Structure conditions.

Each participant learned about two symptoms that were chosen at random, one from a set of meaningful symptoms ("purple spots," "low fluid levels," "cold body temperature") and one from a set of "blank" symptoms ("itchy flippets," "swollen niffles," and "sore mintels"; see Lombrozo [35]). For ease of presentation, we use purple spots and itchy flippets as sample symptoms throughout the chapter.

In the Chain-Structure condition, there were two diseases, Hummel's disease and Tritchet's disease, that could cause these symptoms under some conditions. Specifically, participants read the following information:

> Tritchet's disease always causes itchy flippets and purple spots. One of several ways to contract Tritchet's disease is to first develop Hummel's disease, which causes Tritchet's disease. Aliens can also develop Tritchet's disease independently of having Hummel's disease. Nothing else is known to cause itchy flippets and purple spots, i.e. only aliens who have Tritchet's disease develop itchy flippets and purple spots.

In the Diamond-Structure condition, there were three diseases, Hummel's disease, Tritchet's disease, and Morad's disease. Participants in the Diamond-Structure condition read the following information:

> Morad's disease and Tritchet's disease together always cause itchy flippets and purple spots. If either disease is not present, neither symptom will occur.
>
> One of several ways to contract Tritchet's disease and Morad's disease is to first develop Hummel's disease, which causes both Tritchet's disease and Morad's disease. Hummel's can only cause both of these diseases or neither of them. It will never cause *just* Morad's disease or *just* Tritchet's disease.
>
> Aliens can also develop Tritchet's disease and/or Morad's diseaseindependently of having Hummel's disease.
>
> Nothing else is known to cause itchy flippets and purple spots, i.e. only aliens who have Tritchet's *and* Morad's disease develop itchy flippets and purple spots.

Explanation Choice. Participants in both conditions were told that a particular alien, "Treda," was suffering from the two symptoms. They were asked to choose what they thought was the "most satisfying explanation" for the symptoms from a set of three explanations: the proximal-choice explanation (which included only proximate causes of the symptoms), the complete-choice explanation (which included all the causes they learned about), and an *unknown cause* explanation (see Table 2.1). The order in which these explanations appeared was independently, randomly sampled from a uniform distribution over all possible orderings for each participant.

Explanation Choice Justifications. After indicating their explanation choice, participants were asked: "Why did you choose this explanation?" and could type a few sentences in a text box. We call this the *justification* of their explanation choice.

Reading Checks. Throughout the experiment, participants were asked a series of questions probing whether they accurately understood the causal information presented to them and ensuring they were reading the scenario closely. For example, in the Chain-Structure condition participants were asked whether it is possible to develop Tritchet's disease without having Hummel's disease (the answer is "yes"). If participants failed any reading checks their data were excluded from analyses. The full set of reading checks and exclusion criteria can be found in section A.2 along with the proportion of participants failing reading checks across all experiments.

Table 2.1: Sample questions from Experiment 1. Participants were randomly assigned to either the Chain-Structure condition or the Diamond-Structure condition; the explanation labels (e.g., complete-choice) were not presented to participants.

Explanation Choices, Prompts, and Response Options		
What do you think is the most satisfying explanation for the symptoms Treda is exhibiting?		
	Chain-Structure	Diamond-Structure
Complete-choice	Treda has Hummel's disease, which caused Tritchet's disease, which caused the itchy flippets and purple spots.	Treda has Hummel's disease, which caused Tritchet's disease and Morad's disease, which together caused the itchy flippets and purple spots.
Proximal-choice	Treda does not have Hummel's disease, and independently developed Tritchet's disease, which caused the itchy flippets and purple spots.	Treda does not have Hummel's disease, and independently developed Tritchet's disease and Morad's disease, which together caused the itchy flippets and purple spots.
unknown	Treda developed itchy flippets and purple spots but has neither of the aforementioned diseases.	Treda developed itchy flippets and purple spots but has none of the aforementioned diseases.

2.2.2 Results

2.2.2.1 Explanation Choices.

No participants selected the *unknown cause* explanation. As a result, a percentage of participants selecting the complete-choice (e.g., 80%) implies that the remaining participants (e.g., 20%) selected the proximal-choice. Overall, participants selected the complete-choice 44% of the time in Chain-Structure, and 83% of the time in Diamond-Structure. We analysed responses with χ^2 tests.

Participants selected the proximal-choice and the complete-choice about equally often in the Chain-Structure condition, $\chi^2(1) = 0.22, p > 0.5$, but selected the complete-choice significantly more often than the proximal-choice in the Diamond-Structure condition, $\chi^2(1) = 8.00, p < 0.01$. Reponses across the two Causal Structure conditions additionally differed significantly from each other, $\chi^2(1) = 5.89, p < 0.05$, with the complete-choice chosen more often in Diamond-Structure than in Chain-Structure. These findings are consistent with the predictions of root simplicity, but not with the predictions of node simplicity.

2.2.2.2 Explanation Choice Justifications.

Three coders classified all participants' justifications for their explanation choices into one of four coding categories: "simplicity," "probability," "misunderstood," and "other." Justifications that explicitly appealed to simplicity, complexity, or the number of causes included in the explanation were coded as "simplicity." Justifications that referred to one of the options as being more "probable" or "likely" than the others were classified as "probability." Explanations that suggested the participant misunderstood some aspect of the experiment were classified as "misunderstood," and participants whose explanations fell into this category were excluded from additional analyses. For example, a participant would be classified as "misunderstood" in the Chain-Structure condition if she indicated that Treda must have Tritchet's disease *and* Hummel's disease because that is the *only* way to develop the symptoms. Finally, justifications that did not fall into one of the previous designations were classified as "other." Many of these restated the explanation choice (e.g., "Treda had Tritchet's disease and Hummel's disease which caused itchy flippets and purple spots"), or provided a response that appealed to neither simplicity nor probability, such as "it's what I remember reading from the paragraph," or claiming that it made most sense. Disagreements between coders were resolved in favour of the majority, with rare three-way ties resolved through discussion (Fleiss $\kappa = 0.63, z = 13.19, p <$

10^{-4}).

Overall, 11% of participants justified their choice by appeal to Simplicity, 33% by appeal to Probability, 0% were classified as misunderstood, and the remainder, 56%, fell under Other. The distribution of justifications did not vary as a function of Causal Structure; in fact, the frequencies of response types were identical across the two conditions. Of the small number of justifications that did appeal to simplicity ($N = 4$), two were used to support the proximal-choice in the Chain-Structure condition, none to support the complete-choice in the Chain-Structure condition, one to support the proximal-choice in the Diamond-Structure, condition, and one to support the complete-choice in the Diamond-Structure condition.

2.2.3 Discussion

The findings from Experiment 1 challenge the predictions of node simplicity but support those of root simplicity. Had participants been selecting explanations according to node simplicity, they should have preferred the proximal-choice (i.e., the explanation with fewer causes) in both conditions. Instead, participants were equally likely to choose the proximal-choice and the complete-choice in Chain-Structure, and significantly *less* likely to choose the proximal-choice in Diamond-Structure. These findings conform to the predictions of root simplicity (i.e., that people will prefer explanations with fewer *unexplained* causes), and thereby suggest that root simplicity better describes people's explanatory preferences than node simplicity, at least in these cases.[§] Worth noting, however, is that only a small minority of participants (11%) explicitly justified their explanation choice by appeal to simplicity, suggesting that root simplicity may not correspond to people's explicit conceptions of simplicity.

2.3 Experiment 2: Simplicity and Probabilistic Data

While Experiment 1 challenges the claim that people choose explanations on the basis of node simplicity, the findings cannot differentiate two possibilities for why judgements were consistent with root simplicity. First, it could be that participants' explanatory preferences were a

[§] It is possible that participants were not interpreting these explanations involving relations between causes and effects as causal explanations, but as inductive or deductive arguments or in virtue of some unifying pattern that participants invoked. It is also possible that participants were not choosing the explanations they thought were good, but rather those that they thought were most communicatively natural. Our methods do not distinguish these possibilities, though the results of Experiments 2–4 are more difficult to explain with a communicative alternative. We thank an anonymous reviewer for pointing out these possibilities.

consequence of evaluating each explanation's root simplicity per se. Second, it could be that participants' preferences did not result *directly* from a preference for root-simpler explanations, but instead from assumptions about the relative probabilities of the complete-choice and the proximal-choice explanations; assumptions that happened to align with root simplicity. For example, participants could have assumed (in Diamond-Structure) that Morad's and Tritchtet's diseases were unlikely to co-occur except in the presence of Hummel's disease, and therefore opted for the complete-choice over the proximal-choice on purely probabilistic grounds (i.e., without any recourse to simplicity per se). In Experiment 2, we address this possibility by providing frequency information to indicate how often different diseases occur together in the population at large. This allows us to flexibly adjust the baseline probability of alternative explanations before soliciting participants' explanation choices.

In the present experiment we did not present participants with isolated base-rates (as in Lombrozo [35]), but instead with frequency information that represented the full joint distribution on diseases. Participants first learned a causal structure (Chain-Structure or Diamond-Structure) and then observed a random sample of aliens from the full population. Each alien's disease status (present/absent) was indicated for all diseases. By varying the disease status of the sampled aliens, we could control the relative probabilities of the proximal-choice and the complete-choice explanations for the target alien's symptoms, which participants were asked to explain as in Experiment 1. In this way participants received the frequency information necessary for assessing the probability of each explanation without being told, explicitly, which explanation was most likely.[¶]

Finally, this design allowed us to investigate the effects of explanation choice on memory for frequency information. After the explanation choice task, participants reported back the number of times they remembered having previously observed each combination of diseases in the alien population. Lombrozo [35] found that some participants who selected simple explanations (specifically, those who selected simple explanations which were unlikely to be true) overestimated the frequency of the disease that figured in the simple explanation. Experiment

[¶] It is worth noting that the central findings from Lombrozo [35] involved two sources of probabilistic uncertainty. On the one hand, participants may have been unsure whether the two diseases in the two-disease condition were probabilistically independent, and therefore whether their joint probability was well approximated by the product of their probabilities. On the other hand, participants' "uncertainty" could have stemmed from a more global tendency to rely on an intuitive evaluation of probability when one has to deal with a complex evaluation of multiple sources of evidence. In Experiment 2, we isolate the role of the latter source of uncertainty by eliminating the first: we present participants with data about the full joint probability distribution for the diseases relevant to Chain-Structure and Diamond-Structure.

2 allowed us to investigate whether this effect would extend to cases in which participants were presented with information about the full joint distributions of diseases. It also provided an additional opportunity to differentiate root and node simplicity: if simplicity drives biases in memory for the frequency of causes invoked in simple explanations, these biases should track the simplicity metric that informs people's explanation choices.

2.3.1 Methods

2.3.1.1 Participants.

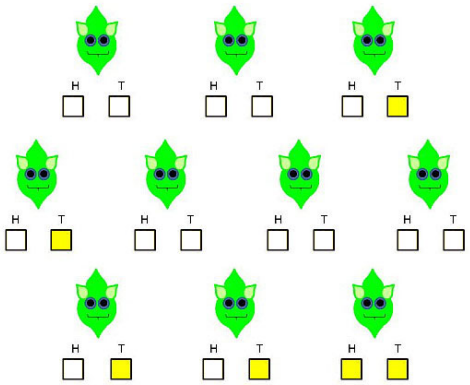
Using Amazon Mechanical Turk, 575 participants were recruited online as in Experiment 1. Of these, 50.6% passed the reading checks described below, leaving 291 participants for analysis.

2.3.1.2 Materials and Procedure.

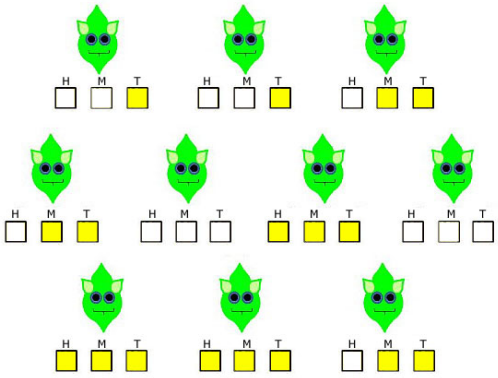
Experiment 2 followed the materials and procedure from Experiment 1 closely. However, before participants were told about Treda and asked to make an explanation choice, they were provided with information about the frequencies at which the different diseases co-occurred. Specifically, participants observed 120 aliens that were described as having been randomly sampled from the population and tested for the presence of each disease. Participants were taught how to interpret images with multiple aliens, with the boxes below each alien indicating the presence or absence of the disease with the corresponding initial letter (see Figure 2.2a and Figure 2.2b). If a box was yellow, that meant that the alien above had the disease indicated by that initial. Otherwise, the alien did not have that disease. Participants were tested to ensure that they understood the representation system as part of our larger set of reading checks.

Participants were told that "the particular incidence rates of these diseases are unknown," but that "to address this issue, the hospital you work in is running diagnostic tests on a random sample of the population" (for complete instructions, see section A.1). The aliens were then presented in 12 groups of 10, with each group appearing for 3.5 seconds between 2-second breaks. In pilot testing we confirmed that these intervals allowed participants to view all aliens while discouraging explicit counting.

We varied the actual frequency information that participants viewed across five between-subjects conditions, the 3:1, 2:1, 1:1, 1:2, and 1:3 conditions, named for the corresponding ratios of the degree to which the evidence supports choosing the proximal-choice versus complete-choice (all frequency counts can be seen in Table 2.2). To compute these *support ratios*, we



(a) Example of a group of 10 aliens from Chain-Structure.



(b) Example of a group of 10 aliens from Diamond-Structure.

Figure 2.2: Yellow boxes indicate the presence of a disease. For example, the top right alien in Figure 2.2b has Tritchet's disease (T) and Morad's disease (M), but not Hummel's disease (H).

defined the probability of an explanation as the percentage of times that the exact pattern of diseases corresponding to that explanation appeared in the data that participants observed. For example, in the Chain-Structure in the 3:1 condition, the *support ratio* is:

$$P(\text{proximal}|\text{data}) : P(\text{complete}|\text{data}) = P(\neg H, T|\mathbf{D}_{3:1}, S) : P(H, T|\mathbf{D}_{3:1}, S) = 3 : 1, \tag{2.1}$$

where

- H and t mean that Treda has Hummel's disease and Tritchet's diseases,
- $\neg$ is the negation operator,
- $\mathbf{D}_{3:1}$ is the frequency data from the 3:1 condition,
- and S indicates the presence of the observed symptoms.

Analogously, in Diamond-Structure the *support ratio* would be

$$P(\neg H, T, M|\mathbf{D_{3:1}}, S) : P(H, T, M|\mathbf{D_{3:1}}, S) = 3 : 1. \tag{2.2}$$

The frequencies in Table 2.2b and Table 2.2a were chosen such that the number of cases supporting the proximal-choice versus complete-choice corresponded to the *support ratio* appropriate for each condition, and for Diamond-Structure, so that the frequencies of M and T were equally likely and approximately conditionally independent given $\neg H$ (to avoid inadvertently suggesting that there existed an additional common cause for these diseases' co-occurrence).

Table 2.2: The frequencies with which each disease combination was presented for each support ratio for Experiments 2(Table 2.2a and Table 2.2b) and 3(Table 2.2a). Note that all columns add to 120 (the total sample). "H," "M," and "T" stand for Hummel's, Morad's and Tritchet's diseases, respectively. "¬ " indicates the absence of a disease.

(a) Ratios for Diamond-Structure.

Diamond Structure					
Event types	Frequency				
	3:1	2:1	1:1	1:2	1:3
¬H, ¬M, ¬T	7	17	33	50	57
¬H, M, T	54	36	18	9	6
H, ¬M, ¬T	1	1	1	1	1
H, M, T	18	18	18	18	18
¬H, M, ¬T	20	24	25	21	19
¬H, ¬M, T	20	24	25	21	19
H, M, ¬T	0	0	0	0	0
H, ¬M, T	0	0	0	0	0

(b) Ratios for Chain-Structure.

Chain Structure					
Event types	Frequency				
	3:1	2:1	1:1	1:2	1:3
¬H, ¬T	47	65	83	92	95
¬H, T	54	36	18	9	6
H, ¬T	1	1	1	1	1
H,T	18	18	18	18	18

Explanation Choice. As in Experiment 1, participants were asked to identify the most satisfying explanation for Treda's two symptoms.

Explanation Choice Justification. Also as in Experiment 1, we asked participants to justify their choice in a free-response format.

Estimated Frequency Counts. Participants were told that they originally observed 120 aliens and asked to indicate how many of these observed aliens belonged to each diagnostic option (presented with its corresponding image), with four possible disease combinations in Chain-Structure (e.g., H but not T) and eight in Diamond-Structure (e.g., H, M, and T).

Reading Checks. The reading checks from Experiment 1 were employed again in Experiment 2. In addition, if participants' responses to the frequency estimate question did not add up to 120 (the correct number) or to 100 (implying a probabilistic interpretation of the question, which we renormalized to add up to 120), they were excluded for failing to follow instructions, and because their inclusion would considerably complicate the analysis and interpretation of the data.

2.3.2 Results

2.3.2.1 Explanation Choices.

All participants who passed the reading checks selected either the proximal-choice or complete-choice explanations. To analyze explanation choices, we first computed the logarithm of the *support ratio* (*log-support-ratio*) in each condition. The *log-support-ratio* should account for explanation choices under two assumptions: first, that participants' explanation choices were a function of the true frequency information provided, and not (for example) a preference for node or root simplicity, and second, that participants "probability matched" — that is, that they chose explanations in proportion to their probability of being true, which is a common strategy in many human judgements (cf. Eberhardt and Danks[75]) and was a useful assumption in interpreting the findings from Lombrozo[35] and Bonawitz and Lombrozo[63]. A systematic deviation from the explanation choices predicted by probability matching would therefore suggest that something other than frequency information (e.g., root simplicity) plays a role in explanation choice.

We conducted a regression (a generalized linear model) on explanation choices with three predictors: *log-support-ratio*, a categorical variable designating each participant's structure (Chain-Structure or Diamond-Structure), and an interaction term to assess whether participants used frequency data differently across structures.

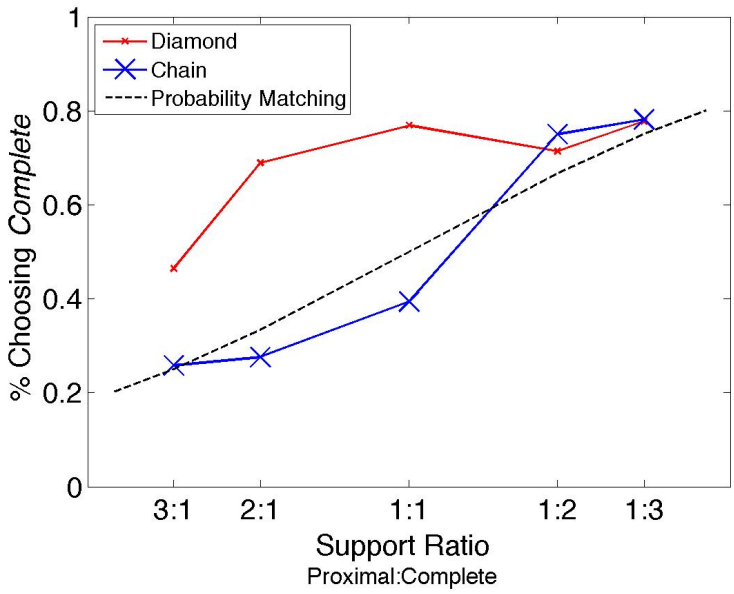


Figure 2.3: Graph of Explanation Choices, % of participants Choosing Complete $\times$ Support Ratio (mapped to the x-axis as log(Y/X) for Y:X, centered at 0 = log(1/1))

The regression, with the proportion of complete-choice selections as the dependent variable, revealed no significant intercept, $t(287) = -0.286, \beta = -0.0515, p > 0.7$, a significant coefficient for *log-support-ratio*, $t(287) = 5.110, \beta = 1.185, p < 10^{-4}$, a significant effect of the categorical variable corresponding to Causal Structure, $t(287) = 3.299, \beta = 0.849, p < 0.001$, and a significant interaction between *log-support-ratio* and Causal Structure, $t(287) = -2.075, \beta = -0.6811, p < 0.05$ (see Figure 2.3[⋆]

The effect of *log-support-ratio* suggests that frequency information had a significant effect on participants' explanation choices, increasing the probability of choosing the complete-choice explanation when it was more frequent in past observations. However, the interaction between *log-support-ratio* and Causal Structure suggests that the influence of frequency information on explanation choices was not equivalent across conditions. We therefore conducted two subsequent regression analyses, treating participants from Chain-Structure and Diamond-Structure independently.

For Chain-Structure, the analysis revealed no significant intercept,$\beta = -0.0515, t(151) =$

[⋆] Technically, this analysis necessitates calculating different interaction effects at each point in question (see Ai and Norton [76]); however, an interaction effect at the intercept is sufficient for our purposes.

$-0.2861, p > 0.7$, but a significant effect of *log-support-ratio*, $t(151) = 5.110, \beta = 1.186 p < 10^{-4}$. The coefficient for *log-support-ratio* did not differ significantly from 1 (95% confidence interval $[0.722, 1.650]$). This analysis suggests that participants' explanation choices in Chain-Structure were well captured by probability matching based on the frequency information that participants received. In other words, the data from Chain-Structure provide no evidence of a preference for either the proximal-choice or complete-choice explanations (above and beyond their frequency), which contrasts with the predictions of node simplicity, but is consistent with those of root simplicity.

For Diamond-Structure, an equivalent analysis revealed a significant intercept, $\beta = 0.797$, $t(136) = 4.224, p < 10^{-4}$, as well as a significant effect of *log-support-ratio*, $\beta = 0.504, t(136) = 2.192, p < 0.05$. In this case, the coefficient for the *log-support-ratio* did differ from 1 (95% confidence interval for $\beta = [0.044, 0.965]$. These results suggest that *log-support-ratio* accounted for some variation in explanation choices, but that participants were significantly more likely to choose the complete-choice explanation than expected on the basis of the frequency information alone. An analysis of the non-zero intercept suggests that participants effectively operated with a prior probability of 0.69 (95% confidence interval for $\beta = [0.603, 0.764]$) favouring the explanation deemed simpler according to root simplicity. This concords with the estimates of the prior probability of a simpler explanation as reported in Lombrozo [35]. These findings also challenge the predictions of node simplicity, but support those of root simplicity.

Deviations from the predictions of *log-support-ratio* in Diamond-Structure were not uniform across support ratios. Post-hoc, one-sample t-tests comparing the proportion of complete-choice explanations for each *log-support-ratio* to the proportion expected from probability-matching revealed significantly higher selection of the complete-choice explanations in the $\mathbf{3:1}, t(27) = 2.619, p < 0.01, \mathbf{2:1}, t(28) = 4.071, p < 10^{-4}$, and $\mathbf{1:1}, t(25) = 2.746, p < 0.01$, cases, but not for $\mathbf{1:2}, t(27) = 0.535, p > 0.5,$ or $\mathbf{1:3}, t(26) = 0.333, p > 0.7$. In other words, participants' explanation choices involved a significant departure from the predictions of probability matching only when frequency information did not favour the root-simpler explanation.

2.3.2.2 Explanation Choice Justifications.

Explanation choice justifications were coded as in Experiment 1. There was moderate agreement amongst the raters (returning all instances of "Misunderstood" to the dataset that were

not excluded for other reasons; Fleiss $\kappa = 0.4415, z = 29.46, p < 10^{-4}$). The distribution of explanation justifications can be found in Table 2.3. We found a significant difference between the overall justification distributions across CausalStructures, $\chi^2(308) = 8.7738, p < 0.05$, with participants more likely to invoke probability inChain-Structure than in Diamond-Structure.

As in Experiment 1, the proportion of justifications that appealed to simplicity was quite small (8%, $N = 25$). Of these, fourteen were used to support the proximal-choice in the Chain-Structure condition, zero to support the complete-choice in the Chain-Structure condition, eight to support the proximal-choice in the Diamond-Structure, condition, and three to support the complete-choice in the Diamond-Structure condition.

	Overall	Chain-Structure	Diamond-Structure
Simplicity:	8.0%	8.9%	7.2%
Probability:	52.4%	58.2%	46%
Other:	33.1%	29.8%	36.6%
Misunderstood:	6.4%	3.1%	9.8%

Table 2.3: Distribution of explanation justifications for Experiment 2.

2.3.2.3 Reported frequencies: Bias for complete-choice over proximal-choice.

The frequency estimates that participants reported at the end of the task were analysed as a function of both the actual frequencies (corresponding to each *log-support-ratio* condition) and participants' individual explanation choices. We considered the extent to which participants overestimated the complete-choice explanation relative to the proximal-choice explanation.

First, we computed the *true difference* between the number of observed cases that corresponded to the proximal-choice explanation and subtracted that from the number of cases corresponding to the complete-choice explanation for a given support ratio. Next, we computed an *estimated difference* by subtracting the number of proximal-choice-consistent cases that a participant estimated having seen from their estimate of the number of complete-choice-consistent cases. Because *estimated difference* should reflect a combination of *true difference* and any biases in memory or reporting, we subtracted the *true difference* from *estimated difference* to create a normalized measure of participants' memory bias for the complete-choice explanation, which we refer to as '*bias*'. A positive value for the *bias* term would result from

overestimating the complete-choice-consistent cases or underestimating the proximal-choice-consistent cases (or both), while a negative value suggests the reverse. A perfect estimate would receive a score of 0 in all frequency conditions.

We analysed *bias* with a linear regression model, using *log-support-ratio* and Causal Structure (Chain-Structure versus Diamond-Structure) as continuous and categorical independent variables, respectively, and *choosing-complete* as a categorical factor. This analysis revealed a non-significant intercept, $t(286) = 0.430, p > 0.6$, a significant effect of *log-support-ratio*, $t(286) = -4.553, p < 0.001$, a significant effect of Causal Structure, $t(286) = 2.525, p < 0.05$, a significant effect of *choosing-complete*, $t(286) = 4.243, p < 10^{-5}$, and a significant interaction between *log-support-ratio* and Causal Structure, t(286) = -3.990, p < 0.001. No other interactions were significant (*p*s > 0.4). Given the interaction between *log-support-ratio* and Causal Structure, and to facilitate the interpretation of these results, we conducted follow-up analyses restricted to each of the Causal Structure conditions and analysed with respect to *log-support-ratio* and *choosing-complete* (see Figure 2.4

In the Chain-Structure condition, the analysis revealed a non-significant intercept, $t(150) = 0.71, p > 0.4$, and significant effects of both *log-support-ratio*, $t(150) = -3.90, p < 0.01$, and *choosing-complete*, $t(150) = 2.40, p < 0.05$. The more strongly the data supported choosing complete the less bias we observed, and those who chose complete were more biased toward complete in their estimates.

In the Diamond-Structure condition, the analysis revealed a marginally-significant intercept, $t(135) = 1.95, p < 0.10$, and significant effects of both *log-support-ratio*, $t(135) = -10.74, p < 10^{-4}$, and *choosing-complete*, $t(135) = 3.71, p < 0.01$. As in Chain Structure, the more strongly the data supported choosing complete the less bias was observed, and those who chose complete were more biased toward complete in their estimates. However, as indicated by the original interaction, *log-support-ratio* had a stronger effect in the Diamond-Structure than in the Chain-Structure.

2.3.3 Discussion

Experiment 2 replicated and extended our findings from Experiment 1. First, participants were no more likely to select the proximal-choice explanation than expected on the basis of probability matching in any condition, challenging the predictions of node simplicity. However, in the Diamond-Structure condition, participants were more likely to select the complete-choice explanation than expected on the basis of probability matching, consistent with the predictions

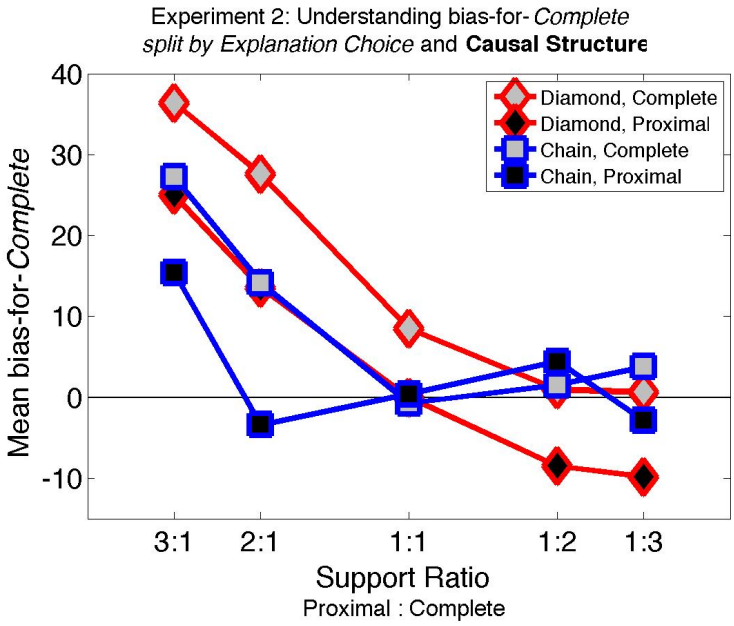


Figure 2.4: Graph of average bias-for-Complete values by Support Ratio, split by Causal Structure and Explanation Choice

of root simplicity. Thus, even when participants are presented with information that (noisily) supported the alternative hypothesis, we see a preference for root simplicity. However, as in Experiment 1, participants rarely justified their explanation choice by explicit appeal to simplicity.

Second, consistent with the findings from Lombrozo [35], participants' explanation choices were a function of *both* simplicity and frequency data. In particular, the proportion of participants selecting the complete-choice explanation was influenced by *log-support-ratio* in *both* Chain-Structure and Diamond-Structure. However, a systematic deviation from the predictions of probability matching emerged in the three *log-support-ratios* involving Diamond-Structure for which the probability information did not warrant the complete-choice: the 3:1, 2:1, and 1:1 conditions.

Third, Experiment 2 revealed a systematic bias in memory: participants who chose the root-simpler explanation sometimes misremembered their observations as more consistent with the root-simpler explanation than they in fact were. This bias emerged in the three conditions for which the evidence did not independently support the root-simpler explanation: the 3:1, 2:1, and 1:1 conditions. These were also the conditions for which explanation choices deviated

from probability matching. In Experiment 3, we consider whether estimation bias was a *consequence* of participants' explanation choices.

2.4 Experiment 3: Simplicity's Effects on Memory

Experiment 2 found that those participants who chose a simple explanation that was not supported by observed data also systematically misremembered the data as more consistent with their explanation than it actually was (see also, Lombrozo [35]. This finding is consistent with the idea that explanation choices can systematically alter memory, but it could also be that systematic distortions in memory have implications for explanation choice (or that both explanation choice and memory for observations have a common cause). Effects on memory due to explanation choices provide converging evidence that a preference for root simplicity is a powerful force in human cognition. Such effects show that the consequences of explanation choices extend beyond explicit explanation evaluation and appear in judgements only indirectly related to explanation.

Here we address these alternatives by varying the order in which participants choose an explanation and are asked to estimate the observed frequency data. Those who estimate before explaining provide a baseline against which we can compare the estimates of those who explain first. If we find that memory distortions are stronger for participants who *explain first* than for those who *estimate first*, this suggests that the act of explaining causes these distortions. If we instead find that distortions are equivalent in both groups, that would suggest that distortions causally contribute to explanation choices and/or that both distortions and explanation choices have a common cause.

2.4.1 Methods

2.4.1.1 Participants.

Three-hundred-eighty-nine participants were recruited via Amazon Mechanical Turk as in Experiments 1–2. Of these, 43.2% passed the reading checks, leaving 168 participants for analysis.

2.4.1.2 *Materials and Procedure.*

The materials and procedure mirrored those from the 3:1, 1:1, and 1:3 Diamond-Structure conditions of Experiment 2, with the following changes. First, we varied the order in which participants were asked to provide their explanation choice and frequency estimates: participants in the *explain-first* condition learned about Treda and indicated an explanation choice before providing frequency estimates (as in Experiment 2). Participants in the *estimate-first* condition were asked to report observed frequencies before they learned about Treda or explained Treda's symptoms.

Second, to ensure that there were equal time delays between observing and reporting frequency data in both ordering conditions, we added an additional explanation choice question which took the place of the alien explanation choice in the *estimate-first* condition. This new question was equivalent to the explanation choice task in terms of time and structure, but irrelevant to the subsequent frequency estimation task. We told participants that a human named Pat was sneezing and asked them for a diagnosis from the following possibilities: "Pat has the flu, which caused her sneezing," "Pat has a cold, which caused her sneezing," and "Pat does not have either of these diseases, her sneezing was caused by something unknown." Because this question is irrelevant to the aims of the study, we do not analyze these data.

2.4.1.3 *Reading Checks.*

Experiment 3 employed the same reading checks as Experiment 2.

2.4.2 Results

2.4.2.1 *Explanation choices.*

Explanation choices replicated those of Experiment 2 (see Figure 2.5), with a significant intercept ($p < 0.001$), a significant effect of log-support-ratio ($p < 0.005$) and no significant effect of task order ($p > 0.5$).[✠]

[✠] As in Experiment 2, we analysed explanation choices using logistic regression with *log-support-ratio* as a predictor for the proportion of participants selecting the complete-choice explanation. However, we additionally included *task-order* (*explain-first* versus *estimate-first*) as a predictor, as well as an interaction term between *log-support-ratio* and *task-order*. This analysis revealed a significant intercept, $t(165) = 3.413, \beta = 0.819, p < 0.001$, as well as a significant coefficient for *log-support-ratio*, $t(165) = 2.815, \beta = 0.536, p < 0.005$. This suggests that participants chose the complete-choice more often than expected on the basis of probability matching, but were additionally sensitive to *log-*

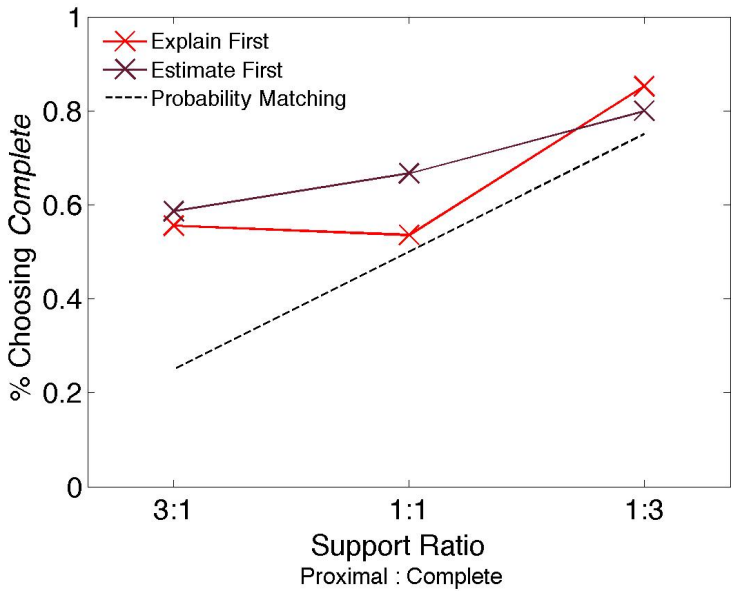


Figure 2.5: Graph of Explanation Choices, % of participants Choosing Complete × Support Ratio (mapped to the x-axis as log(Y/X) for Y:X, centered at 0 = log(1/1))

2.4.2.2 *Explanation choice justifications.*

Justifications were coded as in Experiments 1–2, yielding moderate agreement among coders ($\kappa = 0.577, z = 35.58, p < 10^{-4}$). The justifications distributions differed between the explain-first and the estimate-first conditions, $\chi^2(185) = 7.9078, p < 0.05$, with

support-ratio, with a larger proportion of participants selecting the complete-choice when it was more likely to be true. These findings replicate those from the Diamond-Structure condition in Experiment 2. We did not find significant effects of *task-order*, $t(165) = -0.543p > 0.50$, suggesting that this manipulation did not have a large impact on explanation choices.

Table 2.4: Distribution of explanation justifications for Experiment 3.

	Overall	*Explain-first*	*Estimate-first*
Simplicity:	4.8%	4.2%	5.4%
Probability:	43.6%	49%	38.0%
Other:	40.9%	32.3%	50%
Misunderstood:	10.6%	14.6%	6.5%

participants more likely to provide Other justifications in *estimate-first* (see Table 2.4). As in Experiments 1–2, the proportion of justifications that appealed to simplicity was quite small (4.8%, $N = 9$), with the following distribution across conditions and explanation choices: two were used to support the proximal-choice in the *explain-first* condition, two to support the complete-choice in the *explain-first* condition, three to support the proximal-choice in the *estimate-first* condition, and two to support the complete-choice in the *estimate-first* condition.

2.4.2.3 *Frequency Estimates: Bias for complete-choice over proximal-choice.*

As in Experiment 2, we analysed the magnitude of a bias for the complete-choice. We used *task-order*, *choosing-complete*, *log-support-ratio*, and paired interactions between these variables as predictors of the extent to which participants overestimated the frequency of evidence consistent with the complete-choice over the proximal-choice. As in Experiment 2, the measure of overestimation that we used was "bias," the difference between the estimated difference and the true difference in the number of observations favouring the complete-choice over the proximal-choice.

Most critically, Experiment 3 revealed a significant interaction between *choosing-complete* and *task-order*,$t(162) = 2.961, p < 0.01$, although the effect of *task-order* itself was only marginally significant, $t(162) = -1.929, p < 0.06$. The remaining effects[⊎] of the regression largely replicated those of Experiment 2, and average error across all event types was not influenced by task order[⋉].

[⊎] In Experiment 3, the interaction term between *task-order* and *log-support-ratio* was not a significant predictor, and was thus removed from the analysis, $t(161) = -1.112, p > 0.2$. In the resulting analysis, the intercept was not significant, $t(162) = -0.588, p > 0.5$, suggesting that overall bias did not differ from zero. However, *choosing-complete*,$t(162) = 2.892, p < 0.01$, and *log-support-ratio*, $t(162) = -9.663, p < 10^{-9}$, were significant predictors of bias: participants had a greater bias for the complete-choice in their reported frequencies if they chose the complete-choice as the better explanation or if they were in a *support ratio* condition that favoured the proximal-choice. These findings mirror those from Experiment 2, though here we additionally found an interaction between the effects of *choosing-complete* and *log-support-ratio*, $t(162) = 3.209, p < 0.005$, with the greatest bias favouring the complete-choice-consistent cases among participants who selected the *complete-choice* when it was unlikely to be true.

[⋉] We used a generalized linear model with *task-order*, *choosing-complete*, and *log-support-ratio* as predictors for participants' average absolute error rates across all eight event types. This analysis revealed a significant intercept, $t(164) = 20.317, p < 10^{-9}$, indicating that error was significantly greater than zero, and a significant coefficient for *log-support-ratio*, $t(164) = -4.380, p < 10^{-4}$, indicating that error was greater for conditions that favoured the proximal-choice. Neither *task-order*, $t(164) = -0.471, p > 0.5$, nor *choosing-complete*, $t(164) = -1.189p > 0.2$, were significant

To better understand the interaction between *choosing-complete* and *task-order*, we performed separate analyses for each *support ratio* condition. For the 1:3 condition, in which the frequency evidence supported the complete-choice over the proximal-choice, we found a significant intercept, $t(53) = -3.091, p < 0.01$, a significant effect of *choosing-complete,* $t(53) = 2.691, p < 0.01$, and no significant effect of *task-order (explain-first)*, $t(51) = -0.420, p > 0.6$. Participants had an overall bias for the complete-choice, and had a larger bias if they in fact chose the complete-choice. However, there was no interaction between *choosing-complete* and *task-order* (*explain-first*), $t(53) = -0.245, p > 0.8$, suggesting that in the **1:3** case, where the evidence was sufficient to justify the complete-choice, the bias that emerged was not a consequence of committing to the complete-choice in the explanation task.

In the remaining two *support ratio* conditions, 1:1 and 3:1, the evidence did not favour the complete-choice, and we would therefore anticipate a greater role for explicit explanation choices on frequency estimation, as found in Experiment 2. Consistent with this prediction, in both the 1:1 and 3:1 conditions we found significant intercepts, $t(51) = -3.670, p < 0.001$, and $t(52) = 12.110, p < 10^{-4}$, and significant interaction effects between *choosing-complete* and *task-order* (*explain-first*, $t(51) = 2.706, p < 0.01$, and $t(52) = 2.284, p < 0.05$. In these *support ratio* conditions, participants exhibited a larger estimation bias if they completed the explanation choice task prior to the frequency estimation task and chose the complete-choice.

These analyses also revealed that in the 1:1 condition, there was a marginal main effect of *choosing-complete*, $t(51) = 1.972, p < 0.1$. There was no main effect of *task-order* (*explain-first*), $t(51) = -0.517, p > 0.5$ on explanation choice. In the 3:1 condition, there was a marginal main effect of *task-order* (*explain-first*), $t(52) = -1.729, p < 0.1$: when the explanation choice task was first, there may have been lower bias among participants who chose the proximal-choice than that for those who chose complete (see Figure 2.6). There was no main effect of *choosing-complete*($p > 0.9$).

2.4.3 Discussion

Experiment 3 replicated key findings from Experiment 2: participants were significantly more likely to choose the complete-choice in Diamond-Structure than predicted on the basis of the support ratios and probability matching, with explanation choices modulated by the actual support ratios. This is consistent with the idea that root simplicity and frequency information

predictors of absolute error.

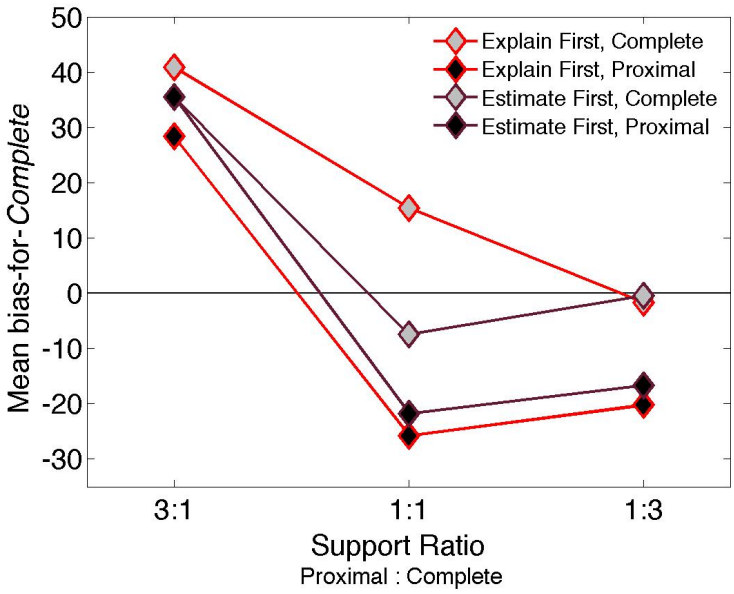


Figure 2.6: Graph of average bias-for-Complete values by Support Ratio, split by Task-Order and Explanation Choice.

jointly inform explanation choices.

Experiment 3 also went beyond Experiment 2 in considering the causal relationship between choosing an explanation and memory. Participants who provided an explanation before estimating frequencies, who chose the complete-choice explanation, and who were in a *support ratio* condition that did not favour complete-choice were most likely to exhibit a large bias for the complete-choice (over the proximal-choice) in their frequency estimates. Notably, these effects were additive: participants were most biased when all three factors co-occurred. These findings not only provide converging evidence for a human preference for root simplicity, but support the idea that its effects extend to judgements beyond the explicit evaluation of explanations.

2.5 Experiment 4: Simplicity and Causal Strength

Why might people favour simpler explanations, especially when doing so appears to have negative consequences for the fidelity of memory (Experiment 3)? Experiment 4 explores one hypothesis about why explanations with few *root* causes, in particular, might be preferred. We

propose that, in general, explanations with fewer root causes provide a useful way to compress information about a causal system for the purposes of memory storage, diagnosis, communication, and intervention. This proposal is related to *Explanation for Export*[74], a hypothesis that suggests explanations are tailored to support predictions and interventions, and so explanations should privilege *exportable* causal information — i.e., information that can be exported from the current situation to support prediction and intervention in novel scenarios (see also Lombrozo[77]). Root causes are prima facie good candidates for exportable causes: they can be used to predict downstream effects, and they make good candidates for interventions intended to have wide-reaching effects (for information theoretic analyses of interventional loci, see *causal information flow* in Ay and Polani[29]; in the context of explanation choice, see Pacer et al.[12]).

If root simplicity is instrumentally valuable – via its relation to effective prediction and intervention – then a preference for root simplicity should be moderated by the degree to which a "root" cause predicts and controls its effects. Specifically, the preference for root simplicity should vary as a function of causal strength, with a stronger preference as the strength of a root cause increases (for more about causal strength see also, Lu et al.[78]). We test this prediction in Experiment 4.

2.5.1 Methods

2.5.1.1 Participants.

Two-hundred-and-five participants were recruited via Amazon Mechanical Turk as in Experiments 1–3. Of these, 57.1% passed the reading checks, leaving 117 participants for analysis.

2.5.1.2 Materials and Procedures.

The materials and procedure were very similar to the 2:1 Diamond-Structure Condition from Experiment 2, and the support ratio was held constant across conditions in Experiment 4. However, the frequency data were varied across three conditions that corresponded to different levels of causal strength between H and M & T: *weak*, *moderate*, and *strong*.

There are several different metrics for causal strength, all of which try to capture the intuition that some causal relationships are stronger than others. Common metrics include probabilistic contrast (Δ_P see Cheng and Novick[79],[80] and causal power ("Power-PC" see Cheng[81]), which once modified to apply to our case scenarios, could be defined as:

$$\Delta P = P\left(M, T|D, H\right) - P(M, T|D, \neg H)\ .$$

$$Power = \frac{\Delta_P}{1 - P(M, T|D, \neg H)}\ ,$$

for positive values of Δ_P, and for negative values, Power-PC is calculated as:

$$Power = \frac{\Delta_P}{P(M, T|D, \neg H)}\ .$$

Across strength conditions, participants received data consistent with a *weak* causal relationship ($\Delta_P \approx .02 and Power - PC \approx .03$), a *moderate* causal relationship ($\Delta_P \approx .28 and Power - PC \approx .45$), or a *strong* causal relationship ($\Delta_P \approx .59 and Power - PC \approx .91$). The final case was identical to the 2:1 Diamond-Structure condition from Experiment 2 (see Table 2.2a for exact frequency counts).

2.5.1.3 Reading Checks.

Experiment 4 involved the same reading checks as Experiment 2.

2.5.2 Results

2.5.2.1 Explanation choices.

We conducted analyses similar to those in Experiment 2. However, because the *support ratio* was held constant while *causal strength* varied, we used the latter as a predictor for explanation choices. Participants were more likely to choose the complete-choice as causal strength increased, whether causal strength was measured using Δ_P, $t(115)$=2.900, $p < 0.005$, or Power-PC $t(115) = 2.895, p < 0.005$.[⊛] The intercepts were not significantly different

[⊛] We reanalysed Experiment 2 using causal strength as a predictor and found largely the same effects. For both Δ_P and Power-PC, causal strength, condition and the intercept were significant (df = 288, $ps < 0.01$). It is worth noting that previous experiments did not manipulate log-support-ratio independently of causal strength; indeed, in all conditions of Experiments 2 and 3, causal strength and log-support-ratio were highly correlated. This results directly from the constraints that we imposed in generating the frequency distributions to represent the different log-support-ratios, namely: having a constant total frequency, a single event in which the root cause occurred and did not in turn cause the proximal disease(s), and (in the Diamond Structure conditions) holding the conditional probabilities of the diseases to be approximately independent given that the root cause was not present (¬H) so as not

from 0, suggesting that there was not a baseline preference for one explanation over another across all conditions in this experiment (ps > 0.9). However, even when the causal strength was weak, participants selected the complete-choice explanation more often than the frequency predicted by probability matching ($\mu = 0.525$, $N = 40$, $z = 2.5715$, $p < 0.05$).

Table 2.5: Frequency data for Experiment 4 causal strength conditions.

Diamond Structure			
Event types	Frequency		
	Strong	Moderate	Weak
¬H, ¬M, ¬T	17	13	9
¬H, M, T	36	36	36
H, ¬M, ¬T	1	9	21
H,M,T	18	18	18
¬H, M, ¬T	24	22	18
¬H, ¬M, T	24	22	18
H, M, ¬T	0	0	0
H, ¬M, T	0	0	0

2.5.2.2 Explanation choice justifications.

Justifications were coded as in Experiments 1–3, with substantial agreement between the three raters ($\kappa = 0.7484$, $z = 24.538$, $p < 10^{-4}$). Overall, justifications invoked simplicity in 1.6% of cases, probability in 52.9%, and other justifications in 40.7%. The remaining 4.9% of participants who passed other reading checks provided explanations that were designated as misunderstood, and were therefore excluded from other analyses. There were two people who justified their explanation choice with reference to simplicity; one who chose complete, and one who chose proximal.

to suggest alternative latent common-cause mechanisms for bringing about the proximal diseases. In light of the systematic correspondences and deviations from the predictions of probability matching (derived from the log support ratios), we think it is unlikely that causal strength alone explains our results in Experiments 2-3.

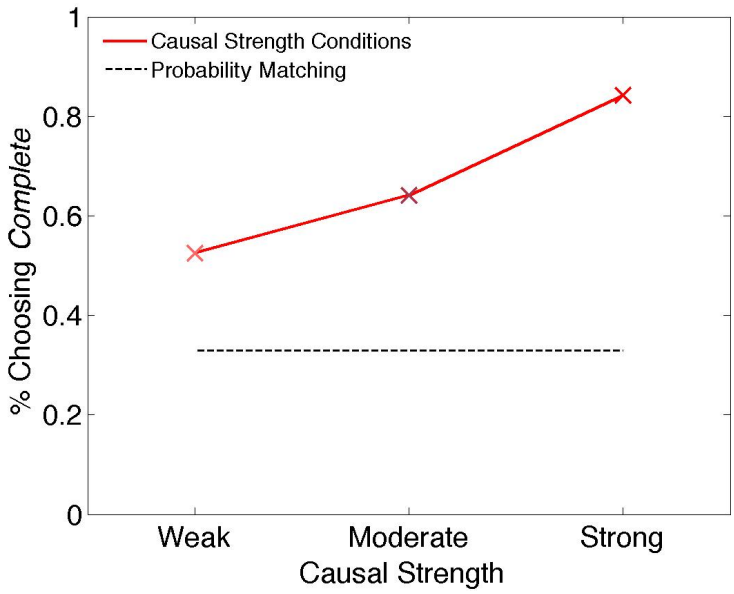


Figure 2.7: Graph of Explanation Choices, % Choosing Complete × Causal Strength.

2.5.2.3 *Reported frequencies: Bias for complete-choice over proximal-choice.*

Each individual's bias for evidence consistent with the complete-choice over the proximal-choice was calculated as in Experiments 2 and 3. Bias was analysed in a regression with *causal strength* (Power-PC) and *choosing-complete* as predictors. This analysis revealed a significant coefficient for *causal strength* $t(114) = 2.844, p < 0.01$, as well as for *choosing-complete*, $t(114) = 4.454, p < 10^{-4}$: participants overestimated the evidence for the complete-choice to a larger degree when the causal strength was greater, and also when they selected the complete-choice. The intercept was not significant, $t(114) = 1.656, p > 0.1$. A parallel analysis using ΔP values instead of Power-PC yielded equivalent results.

2.5.3 Discussion

Experiment 4 varied the causal strength of the relationship between the candidate root cause in the Diamond-Structure (i.e, Hummel's disease) and its two potential effects (i.e., Tritchet's and Morad's diseases). As predicted, we found that as causal strength increased, so too did participants' preference for the complete-choice (the root-simpler explanation), even though the support ratio remained constant at 2:1. This finding is consistent with the idea that a preference

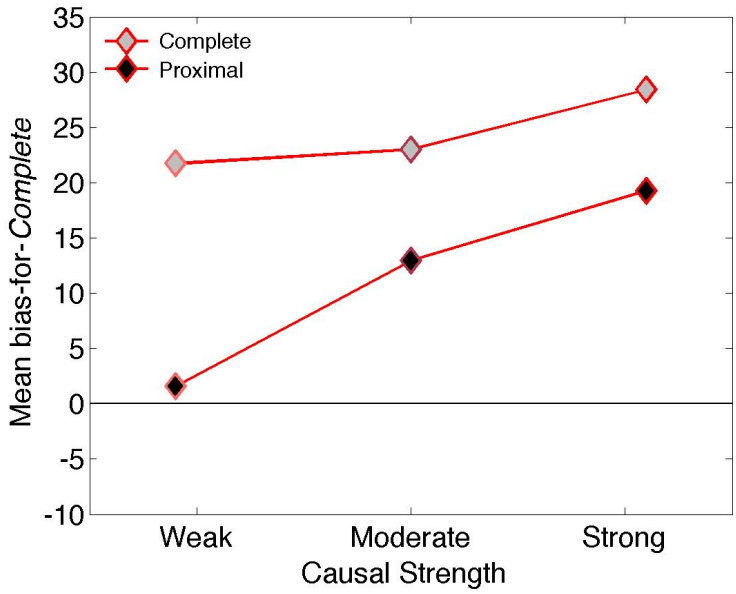


Figure 2.8: Graph of average bias-for-Complete values by Causal Strength, split by Explanation Choice.

for root simplicity derives from the goal of efficiently representing exportable causal information, including causes that effectively predict their effects and support maximally effective and efficient interventions.

2.6 General Discussion

We began by considering four questions about simplicity in explanations and its role in human cognition:

> **Q_1:** What makes an explanation simple?
>
> **Q_2:** How are explanations selected when the simplest explanation is not the one best supported by the data?
>
> **Q_3:** What are the cognitive consequences of a preference for simpler explanations? For example, does the preference bias memory or inference?
>
> **Q_4:** Why do people prefer simpler explanations?

Our findings from Experiments 1 and 2 suggest an answer to $\mathbf{Q_1}$: people's explanatory preferences correspond to root simplicity (i.e., minimizing the number of *unexplained* causes invoked in an explanation), and not to node simplicity (i.e., minimizing the number of total causes invoked in an explanation). Our findings from Experiment 2 additionally provide a partial answer to $\mathbf{Q_2}$: when participants had access to a sample from the full joint probability distribution over diseases, explanatory preferences were a function of both root simplicity and probability.

Experiments 2 and 3 jointly address $\mathbf{Q_3}$, with findings that suggest an influence of explanation on memory for previous observations. Specifically, participants who chose the simpler explanation when the data did not support this choice systematically misreported their observations: they misestimated the rates at which disease combinations occurred in a way that made their explanation choice more likely than it truly was. But they only did this when their chosen explanation was the root-simpler option and was not already supported by the data. Experiment 3 went beyond Experiment 2 and demonstrated that choosing a root-simpler explanation (when it was not independently supported by the data) was a causal factor in subsequent memory distortions.

Finally, Experiment 4 explored $\mathbf{Q_4}$, and found that people's explanatory preferences are more responsive to root simplicity when the root causes are strong. We suggested that people's preference for root-simpler explanations derives from the role explanation plays in generating efficient representations of exportable causal information for prediction and control. The stronger a root cause, the more usefully it fulfils this role. Strong root causes can be used to infer their downstream effects with greater certainty, and strong root causes allow larger or more certain effects from a single intervention. Additionally, we found that people's estimation biases were modulated by the strength of the causal relationship, consistent with the idea that these errors are driven by a preference for root simplicity.

Together, our findings present a unified (if complicated) picture of simplicity and its role in human judgement. We find that root simplicity informs explanatory judgements, is systematically combined with probabilistic information, can alter memory for previous observations, and is especially influential in cases involving strong causal relationships. Interestingly, however, this consistent role for root simplicity in judgements was not reflected in explicit justifications: participants very rarely invoked simplicity or complexity by name, and the small number of such appeals were not restricted to justifications for explanations that were simpler in terms of root simplicity. This suggests that even though root simplicity influences people's judgements, it may not be what people mean when they explicitly justify an explanation with

reference to simplicity.

2.6.1 Relationship to Prior Work

While our findings provide initial answers to $\mathbf{Q}_{1-4}$, they also raise important questions, including their relationship to prior work. For example, we find that simplicity and frequency information jointly influence explanation choices, but how are these two factors combined? Lombrozo [35] argued that simplicity plays a role in determining the prior probability of a hypothesis, but that frequency information influences how the probability assigned to a hypothesis is updated, with a final decision resulting from probability matching to the resulting posterior distribution. The data from Lombrozo [35] suggested a prior for simpler explanations ranging from 68% (Lombrozo [35]: Experiment 3) to 79% (Lombrozo [35]: Experiment 2), and here we find similar results, with priors that range from 68.9% (Experiment 2) to 69.4% (Experiment 3). However, in some cases, we found that participants *underweighted* probability in their final decision (assuming that they were probability matching), potentially because of the format in which the probabilistic information was presented — in Lombrozo [35], it was also the case that frequency information was weighted less heavily when presented in a series of individual cases as opposed to numerical summary values. It could be that data presented sequentially is treated as involving greater uncertainty than numerical summary values.

A second question concerns the way in which explanation affects other judgements, such as probability or frequency estimation. Previous work, reviewed in Koehler [69], has found that prompting people to explain why something could be the case (e.g., a particular team winning a sports tournament) increases the subjective probability that it is (or will be) the case. Our findings from Experiments 2–4 differ in a number of ways. Most notably, we found the largest explanation-induced changes in frequency estimation when the explanation selected was root-simpler *and* when the data themselves did not favour that explanation. In our experiments, explaining itself was not sufficient to strongly alter estimates.

One explanation for the selectivity of our effect is that estimation biases occur as participants try to reconcile two discrepant sources of evidence for the state of the world: their memory for different kinds of observations and their explanatory commitment. Because these are only likely to conflict when an explanatory preference – such as simplicity – draws people to commit to explanations that mismatch their observations, estimation biases are most likely to arise for participants who choose simple explanations when they're unlikely to be true. Future work could investigate these ideas more directly, with an eye towards isolating the effects of

explanation in general from those that arise from specific explanatory preferences.

2.6.2 Limitations and Future Directions

2.6.2.1 Population and materials.

An important limitation to our work stems from the large proportion of participants excluded from analyses, primarily for failing reading comprehension checks. Including such checks is becoming standard practice in research involving data from large on-line populations[82,83], with difficult questions sometimes eliminating nearly 40% of participants[84]. These studies focus on individual exclusion criteria, not sets of criteria used simultaneously, and thus it is hard to compare this past work to our overall exclusion rate. However, none of our individual criteria eliminated anywhere near 40% of participants; the greatest percentage of participants eliminated by a single criterion was 26.5%, with most criteria excluding many fewer (see Supplementary Materials).

Participant exclusions limit the generalizability of our findings to some extent. However, while we suspect that the resulting sample may have been unusually attentive, we have no reason to think that the sample was unrepresentative of the larger population when it comes to explanatory preferences. It is also worth noting that our exclusion criteria were all determined *a priori* and explicitly designed to ensure that participants were correctly understanding the (complex) causal scenarios described; no criteria were modified or added in light of the actual data. Nonetheless, replicating our results with different types of populations is an important step in establishing the generality of our conclusions, as is testing different domains of materials.

2.6.2.2 Individuating causes and causes with internal complexity.

Our analyses have evaluated simplicity with respect to causes that are already individuated and has said nothing about the "complexity" of individual causes. However, both of these assumptions deserve critical scrutiny: simplicity may interact with the individuation of causes, and certainly some causes are more "complex" than others.

These issues potentially arise in Experiment 4, where we suggested that invoking a small number of "strong" root causes allows for more efficient prediction and intervention via more efficient representations of causal systems. If root causes are deterministic causes of their children (which was not the case in any of our studies), then an observation of the root cause is

formally equivalent to observing the cause's children. Faced with that situation, people may re-individuate the causes, representing the deterministic root and its children as a single entity — perhaps as a single cause with more complex internal structure. We expect further study about the role of variable individuation, internal complexity and its relation to preferences for simpler explanations will prove fruitful.

2.6.2.3 Formal metrics of simplicity.

Simplicity has received many formal treatments over the years, and a full story about explanation will assuredly have at least some formal elements. How do our findings relate to these formal approaches, and might our method be adapted to testing formal metrics more directly?

Several metrics consider the number of parameters included in a model and assign models with fewer parameters a higher probability (e.g., Jeffreys [68], Jeffreys and Berger [85], Popper [67], Akaike [50]; or see Baker [26]). Our findings are difficult to reconcile with these accounts without modification. First, such models assume that simplicity is valuable only instrumentally, as a cue to probability, while our results are consistent with a stronger role for simplicity, as participants continued to favour simpler explanations even when evidence unambiguously favoured an alternative. Second, such metrics have typically been concerned only with the *number* of parameters required, not the *values* of those parameters, which Experiment 4 suggests can also modulate preferences for simplicity. Although recent modifications to such metrics have considered the values of parameters, these accounts penalize for large coefficients, i.e., stronger relationships [86], which is the opposite of what we found in Experiment 4, in which stronger causal relationships between the variables resulted in including a higher count of variables.

On penalising large parameters and strong causes. This tendency to disprefer strong relationships between variables (as opposed to numbers of variables) is echoed elsewhere in modern machine learning techniques usually under the term rank- versus trace- or max-norms [87] and regularization [88,89,90,91]. Rank-norms (which penalise weight matrices that have high rank, or number of dimensions needed to factorise), trace-norms (which penalise large weights for a variable averaged across variables) and max-norms (which penalise large weights for a variable maximised across variables) are designed to bound generalisation error in terms of weight matrix complexity in linear classifiers. In this context, one can actually show the line between complexity measures based on strength and those based on dimension are ap-

proximately equivalent; using random projections onto Euclidean half-spaces, low max-norm matrices can be approximated well using low rank matrices[87,92]. In deep learning especially, regularization is often used in order to prevent overfitting, to reduce variance and to bias the parameter space toward useful configurations[89]. It is worth noting though that dropout (creating subnetworks by eliminating variables from contributing to the explanation of particular learning cases during training), is argued to act as a regulariser[91,89] while also encouraging *sparse representations*[90].

Sparse representations are weight settings for which (given a particular input) only a small number of units are expected to have non-zero activations or contributions to the explanation of some learning example. Or, more strongly, "In a good sparse model, there should only be a few highly activated units for any data case."[90]. Small numbers of large activations be seen as consistent with our notion that there is a preference for strong causes in explanations in the final models that are learned, but sparsity (the key claim in the models) also predicts that there are relatively few of these strong causes. This is the intuition behind spike-and-slab priors[93]. This both fits with and is at variance with the literature on people's biases regarding expected parameters in causal induction (which presumably play a role when people learn causal relations from data as they did in Experiments 2–4).

According to some, people's priors sparse (giving few causes weight at all) and strong (giving those causes that have weights large weights) based on theoretical presumptions[78,94]. But empirical estimates of these prior distributions support a prior that favours strong causes, but not sparsity among causes[95]. However, the cases examined in Yeung and Griffiths[95] have relatively few causal variables (2: a potential cause and a background cause) in contrast to the number of parameters that can be present in deep learning models (often $k > 10^6$). At least on the grounds of cognitive capacity, if not for other more computationally grounded reasons, people would *need* to prefer sparse explanations with node-style simplicity (or some way to reduce the number of potential explaining variables) if they were considering networks of those sizes. This suggests that priors that do not express a sparsity preference may only manifest in cases where few variables at play; and indeed when more variables exist as potential competing causes, as in Powell et al.[94], models with priors giving more weight to sparse parameter assignments perform better at modelling human causal strength judgements[ʊ]. Nonetheless, further work is needed to explore the relation between preferences against strong links prevalent machine

Sparse coding and sparse representations more generally are a

[ʊ] In Powell et al.[94], priors with only a preference for strong causes will fail to account for the competition between candidate causes

learning and preferences for strong links in causal induction and explanation.

Kolmogorov Complexity and Algorithmic Information Theory Another approach that is closer in spirit to root simplicity is that exemplified by Kolmogorov Complexity[23] in the field of algorithmic information theory[22], according to which simplicity corresponds to the length of code required to encode something in a Universal Turing Machine. Or (to put it far too simply) the easier it is to compress, the simpler it is. This approach has been advocated most prominently in psychology by Chater (together with computer scientist Vitányi), who has suggested that this notion of simplicity offers a unifying principle for understanding all of cognition[72,25].

While there are clear connections between formal notions of compression (such as Kolmogorov Complexity) and our suggestions in Experiment 4, our own proposal was concerned with efficiently representing a particular kind of information: that which would best support prediction and intervention, and perhaps communication in causal settings. However, Kolmogorov Complexity is an information-theoretic account devoid of causal or interventional information. If we are correct in suggesting that causal information of this sort is relevant for explanation (see also, Pacer et al.[12]), then alternatives to Kolmogorov Complexity that represent causal information (such as *causal information flow*, see Ay and Polani[29]) may need to be developed in order to fully describe these relationships. Exploring the connections between Kolmogorov Complexity and this causally-defined notion of information is a promising direction for future work.

2.6.2.4 Beyond simplicity: other explanatory virtues.

Many other explanatory virtues can (and should) be explored in order to develop a full picture of human explanatory judgements. These include concerns about unification and explanatory scope[71,96], explanatory power[97,98], subsumption[99,100,101], interactions between different "levels" of explanation and general concerns about granularity[31,32,102], among others. Many of these explanatory features have not been analysed in the context of a computational theory of explanation, but we suggest that the paradigms developed here will adapt well to broader exploration, including to cases of non-causal explanation. These additional virtues may conflict with root simplicity; it should be possible to construct cases in which people favour a "root complex" explanation over a root simple one. Our account predicts that this should only occur when other explanatory virtues favour the "root complex" explanation, thereby "outweighing"

the influence of root simplicity. An important open question is how people combine the influences of diverse, potentially conflicting virtues into singular judgements.

2.6.3 Conclusion

By using methods drawn from philosophical, psychological, and statistical toolboxes, we suggest that:

1) *root* simplicity is a better predictor of human behaviour than node simplicity;

2) simplicity trades-off with probability in choosing explanations;

3) choosing and justifying simple explanations can alter memory;

4) and the value of *root* simplicity increases with causal strength.

We unify our findings with a theory of explanation as a process with very specific aims: to inform information-rich representations of causal systems, exportable to other situations in which these representations improve prediction and intervention.

3

Continuous-time Causal Theories*

> Again it came - a throatless, inhuman shriek, sharp and short, very clear and cold. The note itself possessed a minor, metallic quality that he had never heard before. Klausner looked around him, searching instinctively for the source of the noise. The woman next door was the only living thing in sight. He saw her reach down, take a rose stem in the fingers of one hand and snip the stem with a pair of scissors. Again he heard the scream.
>
> It came at the exact moment when the rose stem was cut.…
>
> Klausner shouted "Oh, Mrs Saunders! …Cut another one! Please cut another one quickly!"
>
> She… bent down and snipped another rose.
>
> Again Klausner heard that frightful, throatless shriek in the earphones; again it came at the exact moment the rose stem was cut…
>
> "All right," he said, "that's enough. No more. Please, no more. … I heard them shrieking. Each time you cut one, I heard the cry of pain."
>
> "You might say", he went on, "that a rose bush has no nervous system to feel with, no throat to cry with. You'd be right. It hasn't. Not like ours, anyway. But… how do you know that a rose bush doesn't feel as much pain when someone cuts its stem in two as you would feel if someone cut your wrist off with a garden shears?"
>
> Dahl [103]

Klausner faces an inferential problem: he has built a new machine that amplifies his hearing opening him up to new data. He at first looks around upon hearing the noise; he

* Content from this chapter was originally published in Pacer and Griffiths [13] and Pacer and Griffiths [14] and was co-authored with Tom Griffiths.

knows that a scream he hears must have a causal origin, but the only candidate seems to unrelated. But as the data accrued it suggested that the flowers were source of the screams, inferred from the temporal coincidence of the scream immediately following a stem being severed. He tests the theory by intervening on the system, both indirectly (by social request) and directly (by physical action), and comes to the conclusion that this hypothesis he had never entertained before was true. What data could be so strong as to force him to hypothesise something even he recognises sounds absurd, even mad?

Though the example is fictitious, inferences like it are ubiquitous to human cognition — we are constantly using our causal theories to understand data in terms of an underlying causal structure, but these theories are in flux in order to accommodate data from observation and intervention. It is rare to have one's categories so upturned as to attribute sensation to a new category of things. That said, it is not so ridiculous that we cannot make sense of it; we not only understand his inference but we understand the reasoning for his inference.

Fundamentally, it is the rich temporal information that he uses to justify his startling induction. And even it is only indirectly conveyed (as a summary in written language, not directly experienced) and that conveyance is fictional. But if temporal information ever supports such strong inferences as to alter the set of objects in the world presumed to feel pain, one can only imagine the strength of inferences available to causal theories built of real-time causal knowledge.

The remainder of this chapter addresses that problem: how can we define causal systems using events in continuous-time?

3.1 The ubiquity of rich temporal causal theories

Time is built into causality; the causal theories that shape and power scientific and everyday causal cognition are irreducibly temporal. Given the massive success seen across the sciences, we can confidently say that people have actual causal knowledge about these domains. And the success of science must be based in the capacities of the human mind more generally, so people must be capable of representing entities, properties, relations, events, states, and data defined in terms of time.

However, much of the work on causal induction — particularly work with a formal flavour, has ignored the role of temporal information that is available when one represents data appropriately. I must emphasize that this representation requires a notion of time that can express

both instantaneous events and durative states. Furthermore, those events and states can occur at arbitrary times relative to one another; *there is no grain of experience*.

As a result, if we are to learn from data expressed and experienced over time, our minds need to be able to express events that can actually occur in a continuum. Accordingly, our *models* of the mind need to be able to express continuous-time causal relations, and, furthermore, continuous-time causal theories (CTCTs). In addition to our capacity to represent these events, we need to account for the success in human reasoning about continuous-time causal systems.

The most glaring representational paucity impeding the development of a framework in which continuous-time causal theories can be expressed fully is in the ability to support compositional generative models structured around point processes — infinite sequences of random variables for which most (precisely, almost surely) of the time nothing occurs, but at instantaneous points events occur. This is not claiming that we can perceive infinitesimally short point events but that these points can be useful tools for characterising how we think about events that are effectively treated as occurring at a small proportion of the time. This contrasts with one of the other basic kinds in continuous-time causal theories: states, which have a property of *perseverance* where, for measurable periods of time, they hold the same value.

Given this ontological space for representing events and states – along with assumptions about how to map these events and states to particular sets of observable phenomenon – we have what we need to face the inductive aspect of causal theories. In particular, we can infer the form of the relationship between these events or states (e.g., whether a medicine generates or prevents some side effect) and the structure of relations between different entities/processes (e.g., which medicines affect which other medicines are likely to be taken in the future as a result of taking this medicine).

This chapter will define a formal framework for expressing continuous-time causal theories, particularly attending to the role of continuous-time in the construction of these theories. This framework gives a formal foundation to a series of computational-level, rational analytical models of causal induction, in particular induction of form and structure of causal systems. These models will derive their formal structure and semantics on the basis of experiments on human causal induction; a primary design principle in defining these models is to represent the data conveyed in the experiments with as veridical a structure as possible.

As a result of adhering to the veridicality criterion, models that are similar when viewed from algorithmic and probabilistic perspectives, can nonetheless differ in terms of their theoretical/data representations. In turn that can affect what the model outputs. In some cases, we

can find more than one formal realisation for the same experiment, where multiple theoretical assumptions are compatible with the structure but change the nature of the problem. Some of the times seemingly small changes dramatically alter the results, sometimes apparently significant decisions prove to be irrelevant[†].

On the other hand, the veridicality criterion for experiments that ask multiple questions about the same system, this requires different output models. By doing this I am able to model inference from rates, tabular and real-time data; for functional form and structure induction; with one-shot or repeatable events; with state or point causes; ratings on non-negative scales, integer scales, point-assignment scales, mutually compatible or exclusive scores, and sets of binary questions. Though outperformed by some competing models in traditional cases, my models perform best as the data increases in its potential expressiveness — to the best of my knowledge there are no other published models that can account for some of the experiments I study (and therefore no framework that can account for the variety of cases that CTCTs can).

My single framework accounts well for cases that would traditionally be treated as different kinds of problems. This flexibility only arises because of the attention paid to formal details and the modelling decisions; it is a testament toward the value of paying attention to formal details. However – especially for those not previously acquainted with time-series, point processes, structured probabilistic models, and other technical literatures – these details can be somewhat overwhelming. To allay this, I have bookended each model description with case-studies from the history of medicine that illustrate and ground aspects of the problem in question.

Next I will give a brief overview of some of the history of the notions of causal theories, mechanisms, and causal induction across three disciplines that connect (more or less) with modern cognitive science in fruitful ways: philosophy, statistics and medicine. After this overview, I will review earlier empirical work related to problems of human causal induction focusing on the cognitive science of causal induction with a special interest in work that relates to time and work connected to precise computational models. I then will provide a non-technical overview of some of the central concepts needed to understand the core achievements of rest of the chapter. This should allow a reader with less interest in the formal, computational, or mathematical work to still extract substantial value from the discussion of the experiments. Following that, there are three formal sections: first an introduction to the mathematical and formal background on which the CTCT framework is built, second a desiderata for

[†] See section 3.12 and the discussions of non-zero base-rates, supergraphs and graph filters (minor section 3.12.6.1, and non-zero decay ratesminor section 3.12.7.5

the framework that will cover the topics and issues raised in the historical, empirical and conceptual introductions, and third a description of the CTCT framework in terms of its fulfilment of the desiderata, relation to the previous formal work and extensions that will prove fruitful in the experiments that follow. After that, I describe the application of my framework to build specific models for a variety of experimental results drawn from previous and novel experimental work on human causal induction. These experiments cover many of the topics discussed in earlier sections. I discuss some of the ramifications of this work, how it connects to other work and ways it can be usefully extended. Finally I conclude with a reflection on the relation of time to the study of the mind more generally.

3.2 Influential traditions & theories of causal induction

3.2.1 Philosophy

Induction has long been the hobgoblin of philosophy. For philosophers who see the pursuit of truth as the ultimate aim of philosophy, it is clear why induction appears as bogey. Unlike deduction, where true premises lead to true conclusions, inductive reasoning and arguments make no guarantees. Sceptical responses to the problem have made only made the problem more trenchant, even as they made it more clear.

Induction is unavoidable. No one disagrees with that claim. Any cognitive agent interacting with the real world will be forced to induce as soon as it places its observations into a set of categories that are not singular; that is, any system that applies to more than observation will have to be inductive. Acting in and reasoning about the world without relying on induction is impossible. And people have been extremely successful at acting and reasoning about the world[‡]. Thus, rather than attempting to justify on logical grounds how we *should* induce causal relationships, we can instead ask how, in practice, people *do* induce causes.

That type of question is squarely in the domain of cognitive scientists (Hume [37] even called it a psychological approach). For the more precise projects in this vein, characterising human causal induction relies on carefully defining the problem of causal inference and induction more generally. That sort of care has a long history in the philosophical literature on induction. As a result, theories on what features are relevant to induction and how causal induction *should*

[‡] Any argument of this kind presupposes that past success is predictive of future success, which is the central pillar of the argument Hume [37] builds in his critique of causal induction. As many have pointed out, to perform any analysis (scientific or otherwise), we have no choice but to bite that bullet.

proceed have shaped the questions asked by cognitive scientists. A review of the history of theories of (causal) induction grounds our understanding of why experimental studies and paradigms have taken the form that they have. In particular, it will help explain the persistence of the lacunae my work addresses despite the uninterrupted interest in the topic across the cognitive sciences.

3.2.1.1 Aristotle

Though not traditionally included in the history of the study of causal induction, Aristotle's theory of the four kinds of causes[104] of change provides good grounds for understanding how earlier philosophers thought of "cause" and as a precursor to the internal workings of computational models of causal theories (including those I will introduce as continuous-time causal theories).

Aristotle proposed that all change events can be explained terms of efficient, material, formal and final causes that affect some process. *Efficient* causes are the closest to what I will be treating as causes in my work; they are external events and processes that impinge upon the process that is affected to produce the effected results. *Material* and *formal* causes are, respectively, the properties of the substances involved in the process and the structure and arrangement of the substances in relation to each other. In terms of causal theories, material and formal causes relate to the plausible relations, the form of those relations, and the properties of objects/processes that factor into those relations. *Final* causes are the ends to which the process is "aiming" — this fact is largely unrelated to my purposes, though they will appear again in my discussion (see minor section 3.14.5.2).

3.2.1.2 Hume, Mill and Associationism

Hume[37],[105] is the most influential philosopher on the standard approach to the cognitive science of causal induction, in part through his influence on Mill[38].

Causal considerations aside Hume[37] identified that we can provide neither a deductive nor and inductive justification for our inductive practices. That is, there is no evidence that I could have that would confirm that the sun would rise tomorrow. It is not deductively true, and I cannot induce the claim without presuming the validity of induction on the basis of past evidence, which begs the question. Instead, he suggests that induction is a matter of custom and nothing more. Habits of the mind develop on the basis of repeated instances of the same conjunctions of phenomena, and we proceed accordingly.

Accordingly, he provided a theory of how humans form associations between ideas (which includes ideas that are intended to represent things in the world) on the basis of temporal contiguity, spatial contiguity, similarity and the existence of a causal and effect relationship between the ideas. He notes that cause and effect relationships are the most important in governing our thinking. He then argues that this judgement of cause and effect are equally formed on the basis of the habits of mind, specifically on the basis of a "constant conjunction" of like phenomena in the past, we will continue to expect that the constant conjunction will continue going forward. This constant conjunction required the observation of many occurrences of an event.[§] He rejects the notion of causal inferences on the basis of a single observation, but allows for the possibility that knowledge from similar cases transfers over (which is wise, as he acknowledges too that no two occurrences ever are actually repeated).

Mill [38] took on Hume's [37] challenge and established a set of methods of inference that attempted those insights make more statistically precise. This included the method of agreement (when both cause and effect occur), the method of difference (where neither the cause nor the effect occur), the joint method of agreement and difference (which merely combines the first two), the method of residue (where one has a set of causes for which all of the causes have an effect, but for one and all the effects have been explained, but for one the remaining pair — the residue are causally related), and the method of concomitant variations (where two continuous quantities vary with respect to each other one is either the cause or the effect of the other or they are connected through some other form of causation). As you will see, Mill's [38] methods had a profound impact on later developments in cognitive scientific theories of human causal induction.

Associationism of this sort has echoed throughout different psychological research traditions, impressed by its Empiricist bent. You can see it in theories of conditioning — that is, both classical Pavlov [39] and operant Skinner [106], Ferster and Skinner [40] and operant conditioning. Indeed the behaviourist approach [107] holds as crucial that a scientific enterprise not appeal to hidden and unobservable processes, including mental processes. Accordingly, rules such as Thorndike's [108] law of effect appeal instead to associations between observable events and how they will affect the probability of future events. [109] updated these models by allowing for a notion of a predicted event and errors around that prediction(which itself was fundamentally

[§] It is worth noting that knowing that a phenomenon is a "like" phenomenon to past instances already presumes that you have access to this categorical information. Thus it seems that similarity might be a conceptual precursor to cause and effect, as one could not organise the knowledge necessary for reasoning about cause and effect without first knowing which things counted as multiple instances and therefore a constant conjunction.

unobservable), but even so this work was an associationist approach. Early models of the mind in terms of neural circuitry as a basis for a logical calculus explicitly invoke this notion of learning and causal inference[110]. Associationism echoes in modern approaches to connectionism (especially via back propagation[111]) and reinforcement learning[112] — though by this point the techniques have become immensely more sophisticated and mathematically rigorous than were the original Associationists' accounts.

3.2.2 Statistics and Experiment Design

The earliest work in a statistical vein is likely Mill's[38] methods of inference. The methods offered by statistics to inferring causal relations have grown dramatically in expressiveness, precision, scope and power in the interceding decades.

3.2.2.1 Peirce and Fisher: Randomisation and Analysis of Variance

Peirce[113],[114] was one of the first to explicitly advocate for randomisation in experiment design in order to be able to accurately assess causal influences. Assigning conditions using independent, identically distributed random variables renders the conditions and the interventions they represent as independent of the features they are intended to manipulate, showing this kind of manipulation to be a practical precursor to the notion of intervention used in causal graphical models. And it is worth noting that utility of randomisation was highly contested for nearly half a decade after Peirce's works[115].

Randomisation found an advocate in Fisher[116], whose book on designing experiments successfully demonstrated how randomisation could be used for analysis with statistical inference. Earlier he had shown how to use randomisation to establish causal relationships even when the nature of the intervention meant its effects were nonisolable (such as testing the effects of putting manure on a field in various geometric arrangements)[117]. Also, Fisher[116] popularised analysis of variance, for isolating the source of measurement variance as being due to the result of manipulation or error intrinsic to the system. In addition to becoming a core tool in the statistical practice of psychological science, ANOVA (along with Mill's method) influenced Heider[118] and Kelley[119] as they formulated causal induction in terms of isolating the causes of variation in behaviour as being internal or external.

3.2.2.2 *Karl Pearson: Contingency tables*

Contingency tables are a way of organising categorical multi-variate samples in which each sample is sorted into a group based on its values across a set of discrete categories. The result is a count of how many samples were found exhibiting each combination of categorical values. They were proposed by Pearson [120] to address an absence in the statistical analysis literature: other methods assumed dimensions could be quantified or at least ordered. They were introduced by Pearson [120] as a means of analysing variables that could not be quantified or otherwise ordered. Contingency tables have become crucial formal structures in many cognitive scientists' theories of causal induction [79,81,121,122,61,123].

3.2.2.3 *Wright and Pearl: statistics on graphical models*

One of the most explicit attempts to identify causal relations from statistical data from stochastic processes originated with Wright [124] and his work on path analysis. This work was based on directed acyclic graphs. It can be used to implement most classical statistical models including ANOVA and many more complicated variants. It can be seen as a precursor to modern causal graphical models [15,28]. Both rely on the adherence to there not being any cycles, and aim to be able to isolate the influence of paths between individual variables rather than merely an overall claim about the aggregation of all variables. Causal models like this have played a direct role in or otherwise inspired much of the most recent work on human causal induction [125,126,127,1].

3.2.3 Medicine and Epidemiology

The history of medical and epidemiological inference has played a comparatively smaller role in cognitive science. I believe the cognitive sciences are worse off for this oversight. The insights illuminated by medical history are where the concerns implicit in my work are often best represented. Specifically, the role of time and the inference of hidden mechanism has been crucial for progress in medicine. Time and medicine seem to be inextricably linked.

While many thinkers contributed to our conception of disease, in terms of developing specific criteria for causal induction in medicine, there are a variety of proposed criteria. These criteria differ between inducing different kinds of causal agents such as the effectiveness of a vaccination development procedure, the source of infectious(especially contagious) disease, or occupational disease. While at first this resembles the methodological pluralism found in other sciences' approaches to causal induction, it is not that "anything goes" in all the cases [128].

In fact, likely because of the social strictures placed around medicine – particularly those having to do with the legal, governmental and financial systems (insurance claims with specified standards of treatment, legal suits of malpractice and negligent liability, licensing institutions, multi-year many millions of dollar clinical trials) – few things seem to "go" in medicine once you know the kind of problem you are dealing with. The apparent methodological diversity reflects the diversity in the nature of the causal structures of the subdomains in question.

3.2.3.1 Pasteur and vaccination by successive weakening

Louis Pasteur did not invent vaccination; nor did he perfect the means of developing them. He did develop a sequence of theories regarding how vaccines work, and in tandem with this discovered methods for systematically producing vaccines for a wide variety of diseases from a common method[129]. The first hint of these discovery was a fortunate accident. When a vacation interrupted work on chicken cholera, a month-old culture of bacteria that was intended to be administered fresh did not kill the chickens as was expected. In fact, when a new culture was freshly administered, the chickens were not infected with chicken cholera.

Vaccinations at the time only existed for the diseases that happened to have a less virulent relative that had been discovered to convey immunity. Generating new vaccines artificially had not been accomplished. With his results, Pasteur believed that by weakening the infectious agent, one could artificially generate vaccinations for new diseases. He proceeded to attempt this in a number of ways, including by attempting to pass the disease through a sequence of successively less resilient hosts (e.g., by collecting strains of anthrax that would not kill a horse and giving it to dogs until a strain was found that would not kill the dogs, which was repeated then with rabbits).

This procedure was augmented but its general form to Pasteur and Chamberland[5] became the method of choice for generating vaccinations. This may not seem like causal induction of the standard sort, but when one considers the successive choice among many strains and samples, induction did occur, just by steps rather than leaps and bounds. Furthermore, even with cases where ostensibly data suggested a successful vaccine existed, Pasteur would dismiss attempts to generate vaccines that did not follow this approach[130]. His dismissal rested on the grounds that their method for establishing the weakness of their vaccination strain (and their strength of their testing strain) was insufficient to demonstrate that what they had generated was a proper vaccine. To him, through development in these monotonic sequences could a vaccine be safely found and proved to prevent an appropriately strong version of the disease.

This reluctance to recognise other methods was partially at fault for aggravating the conflict between Pasteur and Robert Koch[131,132].

3.2.3.2 Henle-Koch: establishing infectious agents

Based on earlier work by Henle[133], Koch[131] established the most widely accepted method for identifying the claim that particular entities were biological pathogens; that is, that specific germs were causes of specific diseases[134]. These eventually became reframed as a set of conditions that need to be established, rather than a method per se, given the name the Henle-Koch postulates. The Henle-Koch postulates are as follows (as cited in Evans[134])

1. The parasite occurs every case of the disease in question and under circumstances that can account for the pathological changes and clinical course of the disease.
2. It occurs in no other disease as a fortuitous and nonpathogenic parasite.
3. After being fully isolated from the body and repeatedly grown in pure culture, it can induce the disease anew.

In these, we find echoes of Mill's[38] methods and precursors to later statistical work, but with a predicative logician's flavour for universals and absolutes. These criteria proved difficult to meet once it was acknowledged that there could be asymptomatic carriers. They also were not capable of accounting for chronic, autoimmune, environmental/occupational, or endogenous diseases.

3.2.3.3 Further postulates, toward a unified concept of causation

To account for the great expansion of causal knowledge over the decades that has passed since the proposal of the Henle-Koch postulates, other sets of criteria have been put forth. One set from the epidemiologist Hill[135] focused on identifying environmental and occupational factors (from Evans[134]):

1. Strength of the association
2. Consistency of the association
3. Specificity of association
4. Temporality

5. Biological gradient
6. Plausibility
7. Coherence
8. Experiment
9. Analogy

Some of Hill's [135] criteria seem to closely resemble traditional Empiricist approaches; *strength* and *specificity* under a simple reading could be seen as summary statistics from a contingency table. However, the rest rely on much richer notions of the kinds of causal relations that can exist, an appreciation for the methods of scientific progress, and a drive to unify this claim with the rest of theoretical knowledge. Those sorts of criteria would seem to be irrelevant on simple readings of the Empiricist doctrine. By *consistency of association* has to do the ability of multiple investigations lead by different groups in different places to establish the causal link. *Temporality* refers not only to temporal contiguity but also the duration of the both the cause and the effects, as well as the time-course and order over which the various processes occurred. *Plausibility* refers to how reasonable one finds the attested biological mechanism for linking the cause to the effect(local mechanistic consistency) while *coherence* relates to how well it accords with the rest of our knowledge of natural history, biology and epidemiology (global mechanistic consistency). *Experiment* refers to the ability to affect the disease by experimental intervention, and *analogy* allows for strengthening an argument by citing structural comparisons to other established instances of cause and effect.

These criteria, though, are not specific, and thus are extremely difficult to apply to legal cases, so additional criteria have been posed that are able to be more explicit about what does and does not count as a cause [136]. If one were to want a more strict application for where theories of causality matter as inferential practices (not just as a means of directing an interventional search), the law is the place to expect that causal force from a theory of causality. The modified "Henle-Koch-Evans" postulates (in which Black and Lilienfeld [136] combine the work of Henle [133], Koch [131], Evans [137]) are (as presented in Evans [134]):

1. The prevalence of the disease should be significantly higher in those exposed to the hypothesized cause than in controls not so exposed(the cause may be present in the external environment or as a defect in the host responses).

2. Exposure to the hypothesised cause should be more frequent among those with the disease than in controls without the disease when all other risk factors are held constant.

3. Incidence of the disease should be significantly higher in those exposed to the cause than in those not so exposed, as shown by prospective studies.

4. Temporally, the disease should follow exposure to the hypothesised causative agent with the distribution of incubation periods as a log-normal-shaped curve.

5. A spectrum of host responses should follow exposure to the hypothesised agent along a logical biological gradient from mild to severe.

6. A measurable host response following exposure to the hypothesised cause should have a probability of appearing in those this response before exposure (e.g., antibody, cancer cells) or should increase in magnitude if present before exposure; this response pattern should occur infrequently in persons not so exposed.

7. Experimental reproduction of the disease should occur more frequently in animals or man appropriately exposed to the hypothesised cause than in those not so exposed; this exposure may be deliberate in volunteers, experimentally induced in the laboratory, or demonstrated in a controlled regulation of natural exposure.

8. Elimination or modification of the hypothesised cause or of the vector carrying it should decrease the incidence of the disease (e.g., control of polluted water, removal of tar from cigarettes).

9. Prevention or modification of the host's response on exposure to the hypothesised cause should decrease or eliminate the disease (e.g., immunisation, drugs to lower cholesterol, specific lymphocyte transfer factor in cancer).

10. All of the relationships and findings should make biological and epidemiological sense.

It becomes clear that there are many and varied requirements for a theory of human causal inference that can take all of this into account. Simple counts and proportions without any consideration of structure, time, related theoretical knowledge and a rich ontology for representing a variety of kinds of data will simply not suffice. To have a proper theory of human causal induction we need to take these features into account.

3.3 Empirical background: Phenomena & models

In this section I will review some of the empirical literature on causal induction from the perspective of how the research has addressed (explicitly or implicitly) temporal information.

3.3.1 Limiting features of previous work on causal induction: contingency tables

Earlier work on computational models of causal induction placed strong constraints on the nature of the experiments that were studied. The computations in these cases then lead the experiment which leads the theory. In some cases, the computational abilities of the original theorists(e.g., Hume [37] and Mill [38]) would have played this role, whereas in other cases the inheritance of the theories that were preëmptively limited by carried those constraints as the hindrances of their inheritance. One can see the approach I take as attempting to take the theory and computation to be as rich as needed to capture a variety of experimental methods (especially as regards organising time or collecting data). I then either find extant experiments whose structure helps fill out give structure to the richness of that capacity or design and run the experiments ourselves (as is the case in section 3.11).

The ways previous work has been hampered by adherence to artificial experimental, computational or theoretical strictures illustrate the ways we can exceed those limitations. I first describe the problems in the previous work and then detail a collection of experiments that avoid these limitations and merit further analysis.

3.3.1.1 The 2×2 contingency table

The traditional 2×2 contingency table that is so commonly used in psychological analyses of causation describe the number of times (out of a set of trials) different conjunctions of two variables – a putative cause c and a putative effect e – occur. It consists of four quadrants : (c^+, e^+), (c^+, e^-), (c^-, e^+), and (c^-, e^-), where c^+ indicates that the cause was present, c^- indicates that the cause was absent, e^+ indicates that the effect was present, and e^- indicates that the effect was absent. The number of times the variables take on these values on over all the trials is counted using the counting operator $N(\cdot, \cdot)$ and placed into the appropriate quadrant. Then these counts are used to calculate, for example, functions over their ratios such as Cheng and Novick's [79] $\Delta - P = \frac{N(c^+,e^+)}{N(c^+,e^+)+N(c^+,e^-)} - \frac{N(c^-,e^+)}{N(c^-,e^+)+N(c^-,e^-)} = \frac{N(c^+,e^+)}{N(c^+)} - \frac{N(c^-,e^+)}{N(c^-)}$

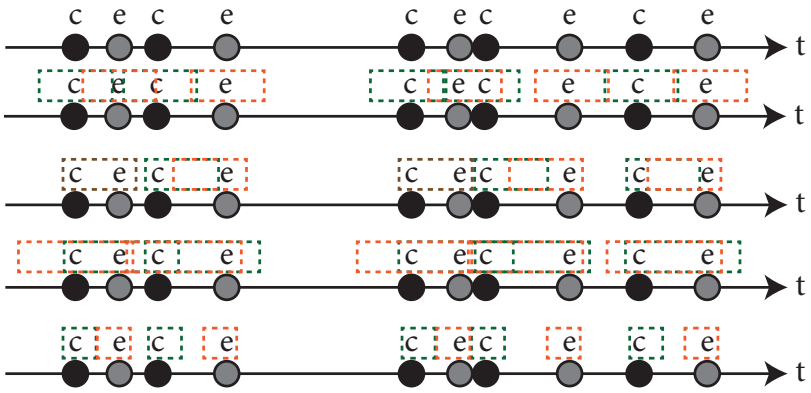


Figure 3.1: Illustration of the problem with binning events in continuous time contingent upon the events. Note, that when counting events this way, no instances of.

or Cheng's [81] power $PC_{gen} = \frac{\Delta P}{1-\frac{N(c^-,e^+)}{N(c^-)}}$ that are meant to act as measures of causal strength (but see Griffiths and Tenenbaum [30]).

Data that occur in continuous time often have causes and effects that do not occur simultaneously, this makes it difficult to encode them in contingency tables in a principled manner. Because of the set of arbitrary decisions that you can make for joining events, a unique encoding of continuous-time data into a 2×2 contingency table is rarely possible.

I am not unique in noticing that there is a deep and profound challenge when it comes to encoding data from the world that is originally embedded in continuous time into contingency tables. Buehner [138] and Hammond and Paynter [139] make similar arguments about the lack of uniqueness (as can be seen in Figure 3.1 ¶. I however seem to be unique among cognitive scientists in my reaction to this recognition that we should abandon theories that rely exclusively on contingency tables and outfit ourselves with new formal tools capable of handling continuous-time directly.

¶ Figure 3.1 illustrates a number of features. One of those features is that that different time scale can result in different views on what evidence there is. That in turn changes which inferences are warranted. It also shows different ways of organising trials around events, and implicitly illustrates the metric problem, where in all of those cases, there is no clear way to identify when there is a case of no cause and no effect.

3.3.1.2 *The representational paucity of contingency tables*

Much of the work on computational models of human causal inference presume the contingency table as the primitive data representation available to the human mind. Furthermore, the table is not a general contingency table but one that has only binary variables as both causes and effects (i.e., it is the 2×2 contingency table described in the previous section). I assert few claims without empirical support of some kind or another, but the following is so basic to my conception of everything that I cannot understand how anyone would disagree with it.

The world is not made of contingency tables; cognition is not built on contingency tables. Even if it were, binary contingency tables would be insufficient. Even if binary contingency tables were sufficient, a two-feature contingency table would miss most of the features that exist in the world. And it is notable that even if one were to generalise it to greater than 2 features, contingency tables do not have a way to distinguish between additive features from substitutive features (those features that can take only one of a variety of values, but will always at least one value will be present)[140,141]. Such a distinction can be crucial for identifying causal relations of different kinds. Indeed without it many of the common causal models based on noisy logical functional forms such as noisy-OR and noisy-AND-NOT would not be definable.

3.3.1.3 *Trial structure*

Contingency tables have built into them the notion of trial structure. That is, to even be able to discuss the conjunction of two variables presumes the existence of a version of time that is discretised into independent chunks that are measured to be observationally equivalent (i.e., each trial is treated as one count and each count is equal to any other count). In most real world settings this is not a tenable assumption; it may be that you can count the number of people who do and do not have a headache at one point in time, but that is not the same thing as tracking the time-course of headaches over that same population.

Worse, looking only at trial structure, and not trials as they are situated in a more general framework of continuous time ignores important aspects of many causal systems. Suppose you were to run the headache study, and gave the experimental group medicine but only tested a year after administering the medicine. We would be unsurprised to discover that the manipulation had little effect, even if the medicine were actually effective. Even though we may not know the mechanism of headaches, we do know a good deal about their frequency and the mechanisms by which most drugs are administered (via diffusion and transport by the blood stream) which provides information about the expected time course of an effect. The amount

of time that has passed during a trial (or even between trials) can matter for the inferences that we make. But analytical methods that focus exclusively on trial *structure* cannot take into account this type of metric information about the amount of continuous-time that has passed.

3.3.2 Trial structure induced from causal events: blickets on trial(s)

One of the most widely celebrated methodological achievements of recent computational cognitive science work is the blicket machine* [142,125,143,144,145,146,147,148]. And this celebration is justly deserved. This mechanism, a toy that can be caused to perform some kind of action (including lighting up, making a noise, &c.) using a remote switch, has allowed extensive exploration of young children's ability to infer causal structure, learn the form of causal relations, and intervene in systematic ways (to name a few of the topics studied). That said, though the blicket studies do not presume that the world is prespliced into evenly spaced trials, they do build into their models and experiments strong assumptions about the causal structure of the world, particularly as regards time.

3.3.2.1 Mackie's [149] malfunctioning machines

To see *how* it does so can be somewhat difficult, and so an example of the relevant distinction can be helpful. To understand this it would help to refer to Mackie's [149] thought experiment of three "coin-slot" machines (i.e., vending machines) K, L, and M. They all claim to supply a bar of a chocolate if if the appropriate amount of money is inserted into the machine, and their internal mechanism is visible. Mackie [149] introduced these to talk about necessary and sufficient causes, we are not strictly speaking concerned with that. What is interesting for my purposes is the way in which they fail; or, rather, the way in which their failures are organised.

K works as you expect; it is deterministic and if you put a coin into the machine you can observe it release a bar of chocolate. If it breaks, you can trust that there will be some account (e.g., some fault in its mechanism) that explains why a coin cause was not given a bar of chocolate in return.

Though instructive for Mackie's [149] purposes, K is somewhat boring for ours (barring an interest in inferring hidden mechanisms). L and M, on the other are indeterministic in usefully

* By the term blicket machine I include cases in which categorisation to a particular label ("blickets") was not the primary objective. Indeed, I include cases in which all that would be said is "The block makes the toy go." as long as the relevant experiment has the event induced trial structure that will be used for the basis of causal inference.

contrasting ways. L is a normal machine in that if you do not put a coin in, nothing will come out. But L is *irreducibly* indeterministic in sense that you may put a coin into the machine and nothing will come out. However, unlike K there is no reason to be given for the case when L fails; it simply does.

M is really the case that we want to examine. Like L, M is irreducibly indeterministic but "its vagaries are opposite to L's"; M violates our usual expectations for vending machines. Occasionally for no reason M will release a chocolate bar when no coin has been put in. It uses the same mechanism as usual, it is just that the mechanism begins acting though no particular cause has been instantiated to produce this event.

We can speak of the relative proportion of times that K and L are successfully activated. We know the number of times that a cause has been initiated, and for each of those causes we know whether the effect occurred. We can directly observe the $N(c^+, e^+)$ and $N(c^+, e^-)$ parts of the 2×2 contingency table. However, we cannot fill the contingency table based on this information. We happen to know that the machine never produces a chocolate bar if no coin is placed into it, but how many times was a coin not placed into it? In the standard case, the period of time during which nothing happened is not easily counted — it is a period of time. One could artificially introduce a count by dividing up that period of time into a number of trials, but then the result for standard contingency table based analyses will heavily depend on that arbitrary choice. On the other hand, one could imagine placing objects other than a coin into the machine and counting the number of times that occurred and how often the chocolate bar was not produced (which would be an equal number of times given the way K and L work). But now we can see that what we are doing in that computation is dividing up time into trials not on the basis of assuming discrete time exists. Rather, this takes it that certain events furnish time with with a trial structure in virtue of their supposed causal effects. The values input to the contingency table type algorithm are not evinced by the non-occurrence but are produced by the events that failed to bring about an occurrence.

The mirror problem can be seen for M, except that the M predicament cannot be solved in the same manner that K and L can be "solved". Why? M has effects that occur in the absence of any cause. That is, we do not need to (and cannot) suppose the occurrence of some event that allows the time to be divided into countable trials so that we can describe the relative proportion of the trials that an effect was and was not present. One *could* assume that these events evince the implicit trial structure of the world, but for aforementioned reasons that is an undesirable step. And it is not that we cannot count anything — the number of times that the effect occurs is perfectly countable. It is the number of times that "no cause" occurs that provides the

trouble.

The only solution to M is to assume that there is some metric over which events occur at some rate. To account for the rate at which effects occur in the absence of causes we invoke a (constant) base-rate. And this same rate based solution could be seen to apply to K and L, just in the case where the base-rate is 0. But then how do we relate this rate of occurrence back to the proportions we obtained when the cause was present? ✠ This is a genuine problem, and much of the work that follows can be seen as a way to address that problem.

To clarify a related but alternative approach, consider that in the cases of K and L note we could imagine instead machines that had a manner of acting that was more akin to M's mode of producing effects in the absence of causes than their own 1-to-1 mapping between cause and effect events. That is, suppose there was a button one pressed and during the time period in which button was pressed it debited your bank account and produced chocolate bars at some rate. If you want that rate could be constant and regular or completely random. The point is that such a mechanism could exist and it would be more akin to the manner in which M, most of the time, experiences a static affair of nothing interrupted by an occasional blip of a candy bar erupting forth. The difference is that in one case existence presses the button, in the other case you are pressing the button.

To foreshadow the approach I would take to solve this problem on the basis of CTCTs, if one imagine the mechanism has some means of ensuring that at most one chocolate bar is released at a time and that a coin insertion (a cause event) can at most produce one chocolate bar (an effect event), then one can reason about both of these types of data in terms of continuous time.

3.3.2.2 Seeing blicket detectors as Mackie's [149] machines

In the case of the studies on blicket detectors, almost universally the researchers have described their data provided to children in terms of the proportions of times that children saw a block (or some other event) go on the machine and the machine activated in some way. The work tacitly assumes that the blicket detectors are K or L type machines. To fill the contingency table they provide events that are cause-like (in that they are also blocks being placed on the

✠ One might respond that events will take some amount of time to occur and then for any one time metric there will be a maximum number of events that could occur and that we should take the proportion in terms of that. However that ignores the fact that one can imagine including many more causal events in a time period than the number of events that could hypothetically occur. If the rate of producing an effect is high enough, this may be what is necessary. Since the proportion needs to be taken out of the number of causal events (not the number of effect events) and because there is no particular limit on the speed at which nothing occurs, any such decision will be fundamentally arbitrary.

machine) but in actuality are supposed to be lacking causal power. Those events then divide up the period of times during which the effect does not occur into a countable number of instances. With that the contingency table is filled and methods that rely on these types of data can proceed. This includes methods based on constraint-based hypothesis tests for learning Bayesian networks as described formally in Scheines et al. [150] and with application to psychological theories of causation in Gopnik et al. [144].

To illustrate this fact even more clearly, we can look at the formal suppositions that Griffiths and Tenenbaum [1] needed to make in order to model these results from blicket type studies. Specifically, the theory that best explained children's judgements Gopnik et al. [125] and Sobel et al. [151] was one in which the children expected a strong exclusive causal relationship between the blickets and the blicket detector, but one with a symmetric error. That is, on any particular trial if a blicket was on the detector it would activate with probability $1 - \epsilon$, but – more importantly – if a nonblicket was on the detector it would activate with probability ϵ. So under this theory, the blicket detector was also a block detector. We need to suppose that because if it were *only* a blicket detector there would be no distinction between the times that a non-blicket block was on it and when nothing was on it at all. And there were such times: when listing the evidence that children perceive (in the case of a system with 2 potential causes A and B), Gopnik et al. [144] mention that the first piece of evidence children observe is that A and B are both absent and that the effect E is absent. It is unfortunate that this data was treated as only a single "trial" (though it was not a part of the data that Griffiths and Tenenbaum [1] included in their model).

A true *blicket* detector (i.e., a machine that detects *only* blickets when placed on top of it) in order to have a representation of noise would need to act like Mackie's [149] M machine. It would need to go off occasionally even when nothing is on it. Otherwise having the error occur exclusively when a block is placed on it indicates that it detects that there is a block on it. It may be that some of the cases where the youngest children studied failed to succeed at the task could be related to a naïve interpretation of the blicket machine as being of M type rather than of L or K. Unfortunately, if we were to want to model this set of studies as though they were true blicket detectors of the M type, we could not do so based on the published literature. The standard protocol for reporting the results of these studies does not involve recording the amount of time that the detector was observed only the relative proportions of the different kinds of causes and their events.

Given the amount of time that children observe most things not being activated by most other things, it may not be odd to attribute a good amount of causal efficacy to those things

that do manage to sometimes activate things. Changing the experimental paradigm to take this background knowledge of the rareness of causal relations could help reconcile a conflict that has been developing in the past decade in the theories expressed by researchers about the role and strength younger children's prior beliefs in learning causal systems. Sobel et al. [151] argue that three-year-olds fail to track category identity base-rates compared to four-year-olds because the four-year-olds were able to acquire a prior probability about categories rapidly. Kushnir and Gopnik [152] argue that three-year-olds fail to follow the evidence because their prior probabilities about causal systems are too strong. Lucas et al. [148] argue that four- and five-year-olds are able to learn a greater variety of causal systems because their prior beliefs are weaker than adults.

We have so much data about what is *not* effective in the world. Nothing happening is the default; processes having no causal effect on one another comprises most of our observations. So when presented with two objects one of which provides a substantial amount of data showing its effectiveness, and the other provides a lesser but still substantial amount of data showing its effectiveness (i.e., that it was effective at all). Perhaps this makes it even more remarkable is that children are capable of learning the relative causal strengths of two different kinds of objects. The research done by those studying causal learning from a computational perspective (in the vein described above) has moved our understanding of the human mind (and our respect for the child's mind) ahead. Research in this vein have already begun looking at how children reason about continuous action sequences and found similar success [153]; that takes this work well beyond what is commonly expressed in contingency tables with uniform trial structure[⊎]. I hardly imagine what the effect would be were this experimental and computational research community to take into account all the data available to the young minds they study.

3.3.3 Paired causes and effects and windows of association

Some of the attempts to generalise beyond contingency tables with trial structures have considered events in continuous time. However, these experimental paradigms were still plagued

[⊎] Though it is not without its own problems. Specifically, the modelling used to explain the human data relied on the assumption that causes needed to be in a contiguous sequence. It is possible that children could reason about delayed effects in a way similar to that described. I believe their model would produce the same results if they did use the appropriately larger hypothesis space. However, doing so may be a waste of effort in comparison to a experimental programme that seeks to address the question of children's causal inferences about long range dependencies more directly

by negative influences that stem from the modelling assumptions. In particular, this work was heavily influenced by descendants of Hume's associationist psychology in the Rescorla-Wagner associationist learning rules[109]. The way conversion from event sequences in continuous time into contingency tables occurs usually relies on a "window of association" in which an individual stimulus can be coördinated with a individual response. This approach has a number of problems.

Time windows can change the inferences you make because they can change whether a particular set of data is an instance of any particular entry in a standard 2×2 contingency table. There are often many ways to do so. Though we are discretising time, it is a matter of convenience, the window of association approach does not require that time be discrete. Accordingly, it will often make sense to establish your time window on the occurrence of an event (see Figure 3.1), but then which events are you to establish it in relation to. For causes and effects are we to know ahead of time whether the time window is to begin before or after each kind of event? Are they to be symmetric around the events? If you make them long enough, then it becomes unclear how you are to count, because one interval could contain several cause events or several effect events in any arbitrary order. What inferences do those merit? The difficulty of such questions play into why many psychological experiments attempt to isolate trials from one another, even in settings where learning (and therefore cross trial effects) are supposed to occur[154]. If, on the other hand, you make the window short enough, no causes will ever be associated with effects and no effects will ever be associated with antecedent causes.

Even more sophisticated accounts that avoid some of the window-of-association problems still may rely on artificially temporal structure imposed by the experimental paradigms that stem from the concepts imposed by previous theories. Gallistel and Gibbon[155] discuss a generalisation beyond this that relies less on the notion of individual delays and aggregation into contingency tables, and instead focuses on the average ratios between different average delays between reinforcements or in terms of different rates of reinforcement. But even so, when their construction relies on the existence of inter-trial delays and the trial durations though they are considering time as a metric, and not ignoring the period of time that elapses between the events, they still are making the assumption that it is possible to achieve a one-to-one mapping between causes and effects (or stimuli and responses).

3.3.4 Human induction: contingency beyond contingency tables

We can gain a great deal by stepping out from the tyranny imposed by contingency tables. One way to escape contingency tables is to modify the kind of data in question by transforming it into a continuous quantity. While work along these lines has been productive (e.g., that of Pacer and Griffiths [13]), I will focus on ways of changing the causal system that focus on how time is represented. In almost all cases, one can frame the underlying data types as "binary" in that they are statements about whether, when and how often events occur.

3.3.4.1 Causal induction from rates

Griffiths and Tenenbaum [30] showed that people are capable of reasoning about causes that increase the rate at which events occur over continuous time, and their judgements are in close accordance with the predictions of a computational model engaging in continuous-time causal inference. In their experiments, participants observed a series of results that they were told came from physics experiments studying whether different electrical fields cause different radioactive compounds to emit more or fewer particles (the compound always released particles at some rate). For each "experiment", participants were told how many particles were emitted during one minute when the electrical field was on and one minute when the field was off. Participants then indicated the degree to which they endorse the claim that the field caused the compounds to emit more particles on a scale of 0 (the field definitely does not cause the compound to decay) to 100 (the field definitely does cause the compound to decay).

Pacer and Griffiths [156] looked at a structurally similar problem for which people inferring preventative causes, where the scale ranged from 0 (the field does not prevent the compound from decaying) to 100 (the field definitely does prevent the compound from decaying). I report this work as one of the more simple cases of CTCTs in section 3.9.

3.3.4.2 Cans distributed in space exploding distributed in time

In Griffiths and Tenenbaum [1] participants are asked to consider a series of four cans full of volatile fluid that each explode at some time. Using their theory-based causal inference model (which can be glossed as a defining a distribution over graphical models) they consider two kinds of hypotheses, the cans exploded because of some underlying perturbation (i.e.,a latent variable) or the explosion of another (nearby) can caused the explosions. To define the time the cans exploded, they simply stated that this was the first arrival of a non-homogeneous pois-

son process whose intensity function $\lambda(t)$ was defined by the underlying graphical structure (the number and relations of latent variables) and the explosions of nearby cans. Importantly, this model can accurately predict people's judgements about the probability that there was a common latent variable causing the explosions based on the number of cans that exploded simultaneously.

The techniques used are related to those I rely on to capture the notion of one-shot events, that is events that can occur once and never again in the history of the model. Once a can explodes it cannot explode again. I give a much more in-depth exploration of how to represent and model one-shot causes in subsection 3.7.8. I will use the approach I develop as part of my models of Greville and Buehner [2] and Lagnado and Sloman [6].

3.3.4.3 Temporal data expressed in tables

One of the features of time that make it such an interesting domain are that events occurring in time can be ordered. Tabular displays are one way of showing the occurrence of different kinds of events across a sequence of time intervals. If one wishes to abandon the regular tabular structure one could represent point events and not just occurrences within an interval of time. Sticking with the regular structure of a table though still allows a good amount of expressivity. You can represent events of different types and whether they occurred in an interval as in Hagmayer and Waldmann [157] or multiple instances of the same type of event and whether they occurred in an interval or not Greville and Buehner [2]. Though I am having difficulty finding examples of it, there should also be a means of combining rate type information about the number of times something occurred within an interval as static rendition of a realisation of the trials shown in Figure 3.2 or how rates like in Griffiths and Tenenbaum [30] could change over time.

Greville and Buehner [2] demonstrated that the temporal distribution of event occurrences will alter people's causal judgements, even if the relative frequencies of the occurrence of the effect in the presence or absence of a cause (i.e., ΔP) is held constant. On the view that all that matters is the relative frequency at which binary events occur in different samples (e.g., the contingency table assumption), this kind of information should be useless. Their purpose was to show that "temporal regularity" influences people's judgements above and beyond mere contingency information. Their experiments used a tabular format to display events that unfolded over five days (split up into five segments of one day each), reporting in which day events occurred. This discretisation allows the use of traditional models of causal inference which infer

causes on the basis of contingency information. This is exemplified by the computational modelling work in Buehner [158], where these exact data are modelled using modifications of the traditional Δ-P and power-pc analyses. Both of those analyses require a notion of a maximum number of "effect-days" which means that as one observed more time, the less distinction there would be between the two cases (as after every bacteria has died both conditions accrue the same number of effect days for every day of observation). Reasoning about these cases should involve working forward to when events occur rather than working backward from the end of observation, as that will not suffer the same effect-day difficulties that this suffers.

3.3.4.4 Inferring hidden mechanism from time and contingency information

We have been viewing temporal data as the underlying representation over which trials are organised and contingency data of the sort that would fit into a contingency table. This is not the only perspective with which to see information about time.

A common approach in cognitive science, which seems at least partially rooted in the supposed conflict between statistical and mechanical causal inference Danks [159] and partially rooted in the overemphasis on contingency tables, is to see temporal and contingency information as fundamentally different kinds of things. In this view temporal information is a "alternative" cue to statistical information, that it influences our judgements (rather than justifies or is incorporated into our judgements), or that it "guides our choices" in what dimensions we attend toHagmayer and Waldmann [157]. The work in Experiment 1 by Lagnado and Sloman [6] is framed in this vein, treating contingency information as primary and temporal information as laid on top of it. Nonetheless, this excellent work has unique features that make it extremely valuable for my purposes.

In Lagnado and Sloman [6] Experiment 1, they ask people to observe 100 trials of computers becoming infected by viruses. In each trial the effect can happen only once. And the participants intervene on the same computer (A) on each trial in order to initiate the sequence of events. From the perspective of the contingency information (i.e., whether on a particular trial a computer is infected by a virus) the causal structure is the same across all four of their conditions.

The important twist on this classical contingency table like experimental set up is that they warn subjects that the times at which the computers manifest their sickness may not accord with the order in which the computers became sick. This cover story allowed Lagnado and Sloman [6] to vary time across the conditions in a realistic fashion. People were able to reason

backwards from what they observed to what which links they believed existed in the underlying causal connection. They found that people's judgements were heavily affected by the timing information. Rather than seeing this as a failure to track the contingencies appropriately, we can see this as an example where people have been given a chance to express their simultaneous dependence on contingency and temporal information.

One reason for this being particularly exciting work for my purposes is that it presents a problem that is truly a case of complete hidden mechanism induction. The cover story implies that the computer has an internal, unobservable state and an external observable state, and that only the unobservable states are causally connected to one another. That means that to infer the causal structures, people had to reason simultaneously about each of the computers as a potential cause *and* as a potential effect. In fact, it allowed people to even postulate the existence of loops, since there was no reason that (when activations spread over time) both routes could be used, even if one is used much more frequently than the other.

If we can build a model capable of capturing people's intuitions, we will have built a model that performs similarly to humans on hidden structure inference. Given that humans are the reigning experts at inducing unobserved causal structures, even in this toy example, this is quite an accomplishment. In section 3.12, I do exactly that.

3.3.4.5 Learning causal structures using real-time stimuli

One of the presumptions that exists essentially across any discussion of causation is that human beings prefer to induce causes with short delays between their occurrence and their effects. This is not the whole story, though because by changing the contextual information and providing relevant theoretical and/or mechanistic grounding you can make people prefer to induce causal relations with greater delays or at least mediate the preference for shorter delays[160,157,161,162]. Indeed some work has suggested predictability may be even more important than a short delay per se[163,164].

Lagnado and Speekenbrink[7] took a different tack at this question. In Lagnado and Speekenbrink's[7] Experiment 2, they asked whether one reason for a preference for short delays might be due to the probability that another event will occur in between the cause and effect variable. That is, all else held to be equal (particularly the underlying rates of events), the longer a delay between a cause and its effect the ore likely it is that some other event will occur within that delay period. They manipulate this directly by having individuals watch videos of "earthquakes" and different kinds of "seismic waves", each of which occur several times in each condition's

video. The conditions vary in terms of the average length of the delay between a cause and an effect and the probability that an intervening event occurs between cause and effect.

The results of this experimental design are a set of videos with precisely encoded point event times and an associated set of mean human judgements about which causes are best supported according to the videos. This is a case where we have maximal information (exact knowledge of when each event occurred for each kind of process). We also have maximally informative data types (where both cause and effect events can occur multiple times). This is an excellent domain in which to apply CTCTs, and so I do in section 3.13.

We have the additional benefit of each animation having only been seen by a single individual, which poses a challenge for classical accounts of experimental methods. Traditional experimental methodologies would say that we cannot have successful inferences because we have only a single sample from any particular stimulus. Nonetheless, CTCTs show how to progress even in this case, because we have a common theory for handling inference in each case, we have effectively performed a principled generalisation akin to making an independent, identically distributed assumption. In the most common case independence is established by having separate trials, which holds here as well. Identicality is normally ensured by having the "same" experimental stimuli, but in this case it holds because of the assumption that the stimuli are being interpreted in terms of the causal theory and thereby are transformed into identical *kinds* of input — what normally is handled by the "repeated" stimulus[⋉] and the uniform identity function is handled as a set of stimuli and a rich CTCT. The CTCT allows making principled generalisations even though no particular stimulus was ever sampled more than once.

3.3.4.6 Responses to past themes in the literature

Overall there seem to be some traditional beliefs from the cognitive science of causal induction that are in need of revision. Fortunately, that process seems to be already underway, but could be greatly accelerated. The work that follows touches on each of these points though I will not emphasise them as much going forward.

Statistics and time are wrongly thought to be somehow fundamentally separable even opposed. Of the two, statistics – particularly, contingency tables – are thought to be primary. All events occur in time, we may represent them as if they did not, but there was a temporal element that was abstracted away at some point to reach that goal. Information is lost when we do this, and if one is not careful an arbitrarily defined false signal can be constructed when trans-

[⋉] Note there are no stimuli that will be exactly the same, even if they can mostly be controlled.

forming continuous-time data into contingency table data. It is safer and more conceptually coherent to work directly in continuous-time and reason backwards. It is not that temporal information or expectations merely guide our choice in reasoning with statistical information, temporal information *is* statistical information.

Real-time causal induction and causal induction from static representations of temporal information are treated as fundamentally different kinds of processes. This stems partially because of the modelling perspectives that apply in one case appear to not apply in another case. Or as Greville and Buehner [2] put it from the computational perspective that favours the static representation of temporal phenomena, " Because real-time paradigms often lack a clear trial structure, calculation of contingency values is impossible." Griffiths and Tenenbaum [30] demonstrate that people make roughly the same inferences about causal induction with rates of events whether they are summarised and presented textually or when presented in real-time. Furthermore they identify

Statistical accounts are treated as incompatible with structural accounts that give weight to prior knowledge. While it may have been true in 2005 when Buehner [138] originally stated "approaches that …address temporal structure are entirely knowledge lean and thus cannot account for …knowledge-mediation effects", this is no longer true. Work on structured probabilistic models like the causal theories of Griffiths and Tenenbaum [1] can represent both prior knowledge and rich temporal structure.

I will now shift to describing the conceptual framework against many of these issues will be able to seen much more starkly.

3.4 Foundational concepts for continuous-time causal theories

3.4.1 Universal, metric, relational, and relative times

If one has a single set of values that are well ordered relative to one another and which acts as the time-domain with respect to which everything occurs, you are using universal time. Furthermore if you are willing to provide a mapping from the real numbers to this sequence of values which allows measuring arbitrary durations, you are using a universal metric time. This is the perspective I will take in this work.

Relational times — where there is some underlying time metric in which the relations are defined — will be of key use, but it is important to keep in mind that these are not *relative* times in the sense of the special theory of relativity.

Relative times would consider that fact that for an event to occur from the perspective of (or having causal consequences on) another process, some time must pass for that information to be transmitted from the perspective of the the affected process. This means that a set of events A, B, C can appear to occur in the order $A \prec B \prec C$ or $C \prec B \prec A$ depending upon the location of the perceiver and the speed at which they are moving relative to the processes that are producing the events being observed. I will not be using this notion of relative time in our work[⊛]. Any instances where I use the term "relative" should be interpreted as "relational", and – unless otherwise specified – it should not be interpreted in the sense of the special theory of relativity.

3.4.2 Entities, Processes, States, and Events

Causal theories are capable of expressing statements about entities, processes, states and states, each of which can have properties associated with them. How these concepts associate with one another (often in ways dependent upon their properties) defines the basic layout of the theory, thus attending to these associations can often be the first step in identifying how to structure an inductive model.

Entities and processes are the organising concepts around which data are defined. Events and states are kinds of data that entities and processes can organise.

An entity can be treated as a "thing" that can be distinguished from other "things" and that persists at least long enough to be observed once (in a trial based context) or for some period of time (in a real-time context). A particular computer is an entity; a particular bacterial culture is an entity. In this sense being a particular entity is a state property of the entity, meaning that entities always have at least one state type property associated with them. Often this property will include a name. An entity will always be defined with reference to state properties, but may have events and processes associated with it.

If something persists, it takes on the same value continuously — identifying names are static properties that continue to apply to an individual. In general, I will use the term "states" to refer to a time series of values where values are maintained for some period of time. When you turn on a light, it would not make sense to turn it "on" unless in some sense once "on" it persisted in being "on". In that case, at all times the light will be "on" or "off" and observing that it is in a state suggests that unless the state changes it will remain in that state. Turning

[⊛] Extending this work to take into account a finite speed for information transmission as well as spatial, kinematic and dynamic considerations is a great project that will need to wait for another time.

the light on may be an event, but you do not continually turn the light on. Contrast that with pressing a button to keep a light on; when you first press the button that is an event but holding the button down is a state.

Processes do not necessarily persist in the same way that entities do . For my purposes, I will be considering processes as "stochastic processes" which is just a way of defining an infinite set of random variables, often with that infinity ranging over time. When a bacterial culture dies, that is the realisation of a process with regards to a particular entity. But processes can range over more than just entities: sequences of earthquakes and seismic waves are processes that do not refer to particular entities outside of the events they manifest.

Often there are many ways to organise sequences of events so as to say that the "same" process brought about each event. Usually this is inferred on the basis of the properties of the events and states associated with the property. For example, while every sneeze is unique, you can see the sneezes from the lifetime of a single person as one process. On the other hand, processes do not need to be associated with particular known entities, the sequence of sneezes in a room with an unknown number of sneezing people as one process (if, for example, you know only the time at which a sneeze occurred and do not know who sneezed when or the spatial location of each sneeze). Processes can also be defined using only the time of occurrence and immediate features associated with the event, as is the case with the kinds of seismic waves discussed in section 3.13 which are identified in a video when concentric circles with a particular colour appear at a particular point and time. There is no underlying "potential ring" entity that the process is tied to, the organising principle is only that the events from the same process appear with the same colour somewhere on the screen.

Most of the time, events do not happen. This is in some ways the definition of an event (at least as thought about in terms of continuous-time point processes). Here I will treat events as though they are instantaneous, i.e., that they have no measurable duration, as a way to formalise this intuition. Even though we cannot observe an instantaneous event with no duration, this treatment of events shows more starkly the role of continuous-time in our theories. Firstly, human causal theories have postulated the existence of instantaneous events — neurons were considered to conduct neural impulses instantaneously until they were demonstrated by Helmholtz in 1848 to have finite velocity[166]ↄ. Therefore, we can think about events

In actuality, taking a cue from Salmon [165], we can understand entities as being processes that – for each entity – consists of all the processes that define an individual entity's features over time.

ↄ This was in contrast to du Bois-Reymond [167], who argued for the infinitesimal nature of the signal. Interestingly, this not long before du Bois-Reymond's [167] brother, du Bois-Reymond was researching infinitesimals and their mathematical foundations (c.f., for a further discussion on theories of continuity

that we acknowledge we are unable to perceive, and such thinking may grant access to properties that would similarly be impossible to perceive. For example, this allows ensuring that no two events happen at the exact same time ♣, meaning that we can say that every event happens before or after any other event. Even if you allow simultaneous events, you still have a partial order. But when you have states, their time periods can overlap (for example, if you consider the time periods in Figure 3.1 as states some of them overlap), meaning that you cannot even say whether states occurring as parts of different processes happened before or after one another.

You can actually see sequences of states as static properties that have events that define their change-points. Seen in this way, there will be events in between each state transition with properties identifying the "original state" and the "new state" as well as the particular time at which the transition occurred. At least, this is straightforwardly definable in terms of continuous-time, but in discrete time it is not clear in which time step the state transition event occurred. It could be joined with the time step of the original state or the time step of the new state ♦, but there is no unique interpretation. By convention in probabilistic models it is usually joined with the new state, because you do not know that the transition occurred until that time step, but this is merely a convention. But then you can define a process that has the value of 0 on the time steps where a value stays the same as the previous value, 1 on those time steps where the value differs from the previous value. But now what I have been calling an event seems to be defined as a series of state values.

If you only consider discrete time it can be challenging to distinguish between events and states, because there is no way to distinguish between an event occurring during a timestep and taking on a state value for that timestep. In discrete time there is a minimal duration, and so our instantaneous events must also occur over that duration. In its simplest form no more than one event can occur within a discrete time chunk as the definition of the discrete time chunk, but you could also count the number of events that occur within time chunk. However, even in that case, you can treat the number of events as a state that happens to manifest for that time chunk, meaning there is still a symmetry between events and states. Continuous time with point events allows escaping this conclusion and obtain an asymmetry between events and

in mathematics around that time Buckley [169].

♣ Technically, stating that no two events happen at the same time requires further assumptions such as orderliness or simplicity in the point process, but I will address that later.

♦ Though it may be appealing, treating it as occurring "between" the original and new states is a notion of time that violates standard accounts of discrete time. If you have discretised the world, there is no time for it to occur between the state values.

states.

3.4.3 Discrete and continuous time

Discrete time treats all processes as occurring over a countable number of "chunks" of time. Most often each chunk is treated as the "equal" to any other chunk, there is a new time chunk at each "tick" of a "perfect clock" organising all the events in question. For a process defined over discrete time it will have a value on every tick that has already occurred. But you could define discrete time differently, for example, dividing your life up into the different periods during which you have been ill or not ill, which will hopefully not be uniform. But at all times steps throughout in your life you are either ill or not ill.

Discrete time has a number of interesting properties. Discrete time spans all of the instances during which values could possibly be observed. There is only one time step sequence that applies to all processes — processes relations at different time scales, but those relations must be defined relative to a integral number that specifies the number of time steps that occurred between. Every time step is either before, after, or is the same as another time step [♥] and no events (or state values) can occur between time steps. Thus an endorsement of discrete time provides a grid by which all events in all processes are to be organised. If you take discrete time as a primitive construction, that is essentially a priori discretisation; other discrete time metrics can supervene on this, but all of their "time-steps" will have to be specified relative to the time-steps of the primitive metric.

Continuous time has no discretisation ahead of time, though it can be discretised if that is wished. This could be pre-gridded (an a priori discretisation), which effectively lays a discrete time measure on top of the continuous measure (see Figure 3.2 and how I integrate over days in section 3.10). This can be seen as the inverse of the transform that is often used to illustrate continuous point processes as limits of Bernoulli processes as in the comparison be-

[♥] I will not address this again, but it is worth noting that much of the philosophy of time would object to this characterisation that only states that times need to succeed one another. More precise theories of temporal order state events in terms of "betweenness"[170,171]. This means to uniquely order a set of events you need at least four events that occur at four different times to be able to say that in $\{A, B, C, D\}$, B is between A and C and that C is between B and D. This arises to avoid paradoxes under more general topologies for time such as cycles which would have the paradoxical conclusion that A was before D and D was before A. These considerations are not central to the arguments of the chapter, and it makes an already confusing explication even more convoluted. Given that I will be working over a time metric (a much stronger assumption than a partial order), these considerations can be safely placed to the side.

tween columns 2 and 3 of Figure 3.2 Though for certain data types this can pose ambiguities that require interpretation. For example, if you are reasoning about repeated events but wish to represent them as a Bernoulli process (an infinite set of Bernoulli random variables) it is unclear what it means if you find you have more than one event occurring within a time bin. And even that construction relies on the ability to encode the occurrence/absence of an event as a Bernoulli variable. But it can be difficult to do this if one has state-type processes defined over the underlying continuous measure, because almost surely the transitions will not align with the prespecified grid. In that case, even if one only has two states(i.e., a binary value) then can be unclear how to assign a state to the time-step during which value changes took place.

Continuous time allows expressing this notion: most of the time, nothing happens; but over time, things happen. This is fortunate. We expect that the ground on which we stand will not flit in and out of existence — it is a persistent support. In some ways perception, inference and action must be built on top of representations that have some persistent qualities. But they also need to be able to account for events (even if they are only change-points in otherwise stable states). Entities or processes will need to persist, but events that involve or constitute the entities or processes can be observed in relation to the otherwise stable background. This kind of normalcy and perturbation information is crucial to inducing workable causal relations in the world. Not all causal induction needs to be understood in terms of perturbations to a normal state, but points in a continuum are particularly amenable to this kind of analysis. Because I treat events as points, any particular occurrence have (measure) 0 probability of happening; with probability (measure) 1 nothing happens at any particular future time. But over a measurable chunk of time, we can expect some number of occurrences at particular points. After they have occurred, they are atomic events (in that they still have measure zero) that have probability 1 of having occurred. After the fact(*a posteriori*), events can play causal roles and act as prospective organising principles for evaluating future events and states.

3.4.3.1 A posteriori discretisation and trials

We say that time is a posteriori discretised when you divide up time based on when an event occurs, after that event occurs. This is the case in Figure 3.1. Usually this also involves an ending event that ends the observation, and a function describes the events and states that occurred during that period of time. The discretisation supervenes on the future states of the basic temporal structure, prospectively defining later events in relation to the discretising event.

One can overlay a posteriori discretisations on either kind of time, where time is discretised

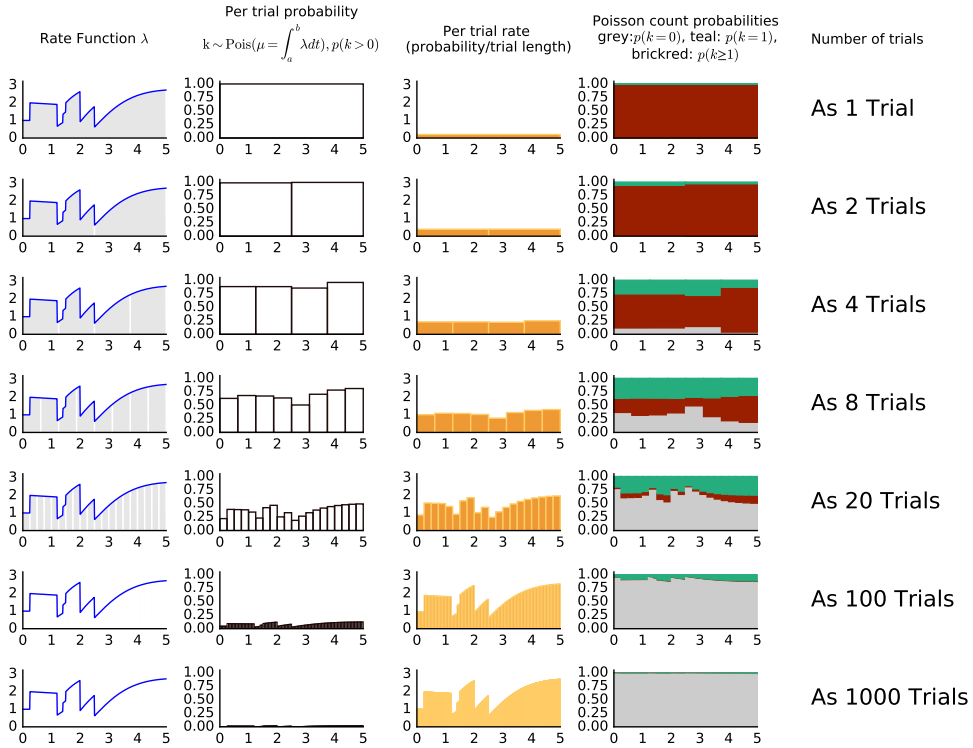


Figure 3.2: This figure shows the primacy and inevitability of rates. You can always recover discrete-time trials and probabilities for whether at least one event occurred by integrating over continuous-time rate function to obtaining the expected number of occurrences according to a Poisson distribution. If it is a binary, discrete-time trial then it will only be able to account for one of those events even if many events occurred. In the first column, we see the rate function divided up into 1, 2, 4, 8, 20, 100, and 1000 trials; one for each row of graphs. The second column shows the probability of at least one event occurring in each trial. The third column shows the rate of at least one event occurring in each trial, which we can compute by dividing the probability of occurrence by the length of the trial. The final column shows the degree to which Bernoulli trials successfully capture occurrence in their approximation; the brickred areas are the probability that a trial will have had more than one event occur (and therefore be tossing away data if analysed by a Bernoulli process), the grey areas are the probability that no event occurred (which suggest that instances of observations being treated as Bernoulli trials are mostly wasted with nonobservations), and the teal areas indicate the probability that exactly one event occurred. Notice how widely the distributions of the number of events can vary depending on how you discretise time(contrast the different areas found in the fourth column and how they vary across trials). Notice also, that as you discretise the rate into smaller and smaller pieces, the probability of no events rises to be nearly 1(column 4), the probability of any event occurring goes to 0(column 2), but the rate of event occurrences approaches original rate function(column 3).

according to some set of events that actually occur. This can be defined in many ways. The previously mentioned division of time into the times you were sick and the times you were not sick are state-type processes that can define an a posteriori discretisation of your life. Or one could count a number of events assumed to be occurring at a constant rate and use that to determine how much time has passed. In fact, this is how the standard definition of a second is defined [172]; that is, a second according to the official standard by of Weights and Measures [173] is the amount of time it takes (on average) for caesium to release $9,192,631,770$ instances of radiation. Aggregated point events become the basis of our time divisions which can be expected to be regular due to the relative number of those events given the underlying stochasticity of their oscillations between two ground states.

However, if the overlay is to be done on a system defined over discrete time, the units of the overlay must be constructed in terms of the integer number of time steps. Thus if we prediscretised the world into "days", we would need to modify the definition of the previous sickness time-scale to fit within days. There may be many ways to do this, for example a new definition would be: divide time into those sequences of days in which you (at any point in the day) were ill and those sequences of days for which you were not ill at any point in the day. Alternatively, you could define it to be sequences of days in which more than half of the day you were ill and sequences of days in which you were ill for less than half of the day. No matter what, the definition will have to ground out in days.

Continuous time allows a posteriori discretisation with arbitrary beginning and ending points, including the ending point of ∞ (in terms of available representations) or whatever the current time is relative to the starting point T (in terms of available data). This could create the same ambiguities seen in discrete time by somehow summarising data that occurs within the time period. But there is no basic granularity for the available data; if you you become ill partway through one day, you just record as an event that time at which you became ill. Though continuous time can be discretised, discrete time cannot be continuised (made continuous) without further assumptions.

You can also plan out a priori a strategy for a posteriori discretisation once you have the data that you planned on receiving. This type of strategy is crucial for developing a clear conception of the relation between theories and data, as they are actually used by practicing scientists and statisticians. The basic idea of a "trial" or a "sample" is that some set of observations can be divided up into discretised summaries of the activity that occurred during that observation period.

"Trials" are crucial to a modern conception of scientific and statistical practice are often

defined in relation to data that is acquired over a series of "trials", which are usually assumed to be independently and identically distributed. That is, we observe the value of a process at some number of time chunks and aggregate that information across those time chunks. Additionally, this requires assuming that the values obtained from each of them does not affect the value obtained from any of the others and that the process gave the same probability of observing values for every trial. Trials of this sort could occur in discrete time: they begin at a time step, persist for some amount of time(possibly only one time step) and then end on a time-step.

Discretising time into trials is a separate activity from treating the world as discrete time. Fortunately for scientists, unlike discrete time, discretising time into trials does not need to span all of time. There can be time periods during which no trials occur. If there is more than one process to be measured, one could have trials that occur within partially overlapping periods of time and nonetheless treat them as having occurred at different times despite themselves being individual discrete units for the purposes of the data they generate. Two individual discrete units of discrete time would need to occur in the "same" time step in order to be atomic for the purposes of the data they generate.

3.4.4 Additive and substitutive features

The variety of time we choose to use as our fundamental representation affects the kinds of features that can easily exist within our theories. These different kinds of features, especially when allowed to be in relation to one another across entities and processes, allow for different varieties of categorical structure even in the case where time is not considered[140]. These differences are amplified if we treat time as fundamentally continuous; there are asymmetries in continuous time that are absent in discrete time. To illustrate this point it will be useful to discuss two kinds of binary features: additive and substitutive binary features.

Additive features are either present or absent. My computer does or does not have an external keyboard. My keys do or do not make sounds when they are struck. A molecule of radium does or does not decay over some period of time. Additive features are fundamentally asymmetric in that they have a state *absence* which requires the existence of base-line values embedded in the theory that describe the object absent any particular features. The notion absence most relevant to my purposes can be illustrated using examples.

Suppose you observed two objects of the same kind, but which had one present additive feature. Now imagine you observe the negation of the additive feature, i.e., you now have two objects of that kind for which both additive features are absent. Were you to be shown one of

these two "base-line" objects and asked to return it to its original state, you would have no way to know how to do so — one base-line object is indistinguishable from another. In fact, any object has an infinitude of potential additive features that happen to be absent ♠ A modified version of negation that negates only present additive features is an identity-losing operation in that after that operation, previously distinguishable objects would cease to be distinguishable — all objects revert to the base-line object. As a result if one were to attempt to negate that again, there would be no way to identify which features were originally present based on the feature values of the once transformed object. This arises because the second negation could return you to any set of the additive features, as they can be present together. That is where the additivity comes from, I can add features without conflicting with the presence of other features; I can strike more than one key at the same time, and it does not eliminate the other key presses.

Substitutive binary features are features that can take on one of two values without any asymmetry between the features. My monitor cannot both represent my text and be on a screensaver at the same time. Substitutive features can have more than two values, for example, my keyboard sends only a single set of signals at one time (even if that signal consists of multiple key-strikes). While there may be an infinite number of values a substitutive feature could take on, in practice they are often limited to a small discrete set. In those cases, there will not be an infinite set of substitutive features that could apply to any entity or process because not all entities or processes will take on either of the values of the substitutive feature. In a sense, you could see substitutive features as built on top of the additive feature of the existence that substitutive feature, whatever its value. Because each feature is defined using mutually exclusive, symmetric values, standard double negation can occur(because of the symmetry) without information loss (because of binary exclusivity). For any set of binary substitutive features the first negation swaps the feature values for their opposite which does not have a guaranteed common interpretation with the opposite value of other substitutive features like "absence" does in the additive case. Then, the second negation returns the losslessly stored information about the original values which can be derived without ambiguity from the opposite values. ❖ Then the second negation recreates the original feature values for all the features. There is

♠ Because there are an infinite number of absent features, theories need to be able to characterise objects as having notably absent additive features in order to use evidence that features are absent

❖ In the n-ary case, to remain lossless, the negation will need a modified interpretation that allows value-sets that encompasses the possibility of any of the other values (as there is no unique opposite to the observed value in the n-ary case). If instead you were to force it to choose a particular value among the value-set, it would lose information about the original value that cannot be restored with a

no necessary "base-line" case that different substitutive features revert to and so no confusion occurs between the substitutive features.

For my purposes, the distinction between these two kinds of features will revolve around a stronger interpretation of this asymmetry in the additive case and exclusivity in the substitutive case.

In continuous time, we define point processes as points against a backdrop which is understood to be mostly comprised of the *absence* of points. That absence consists of an (uncountably) infinite set of potential points that are simply not extant. Multiple point processes can be superimposed upon one another (almost surely) without creating conflict between the values of the processes; you might lose the identity of where each point originated(if no additional information is recorded), but the result is still a valid point process. And similarly the negation of any realised continuous time point process is indistinguishable (in terms of its measure) from the negation of any other realisation over the same space. In these senses, they are analogous to additive processes.

State processes, on the other hand, have persistent values that extend over continuous periods of time. You could interpret a state as an "absent" feature, but that interpretation is not necessary as it is in the point process case. Even if you did so,"absence" could be mapped to either state value arbitrarily. The only constraint on the meaning of "absence" arises out of the roles given to the states in relation to other parts of the causal theory (e.g., see section 3.9), rather than as a definitional feature of the formal structure as it is in point processes. Multiple state processes occurring over the same space cannot be superimposed on one another, because the states' intervals from the different processes will conflict over the continuous intervals in which they hold. One may be able to define such an operation, but doing so requires either losing information in the case of conflict (as only one process' state would be preserved) or re-representing the conjoint processes using higher-dimensional features over the space. Again, this arises because "absence" is absent — there is no guaranteed state that spans the support space that is shared between different features and does not conflict with the values assigned to those different features. As a result, there is no guaranteed way to combine the information from mutually incompatible states.

---

second modified negation, because the original value is not uniquely defined as the opposite of whatever the chosen value is. However, this is a different type of modification and ambiguity than that which is needed and arises in the additive case, where there is an asymmetry inherent in the meaning of the features' values (only present features are negated) and feature identities (not the values) are what is lost under the modified double negation.

3.4.5 Background processes, hidden causes, generators and preventers

When I refer to a background process this is meant to include all the unseen causes that are (by stipulation) unobservable and uniform. That is, it is the "background" because the activity due to the background cause is effectively spontaneous or unattributable to any specific causes. This feature is necessary in some cases where we want to enforce strict conditions (e.g., that causes should always precede their effects) but allow models and simulations otherwise would be treated as impossible to be merely unlikely by accounting for otherwise unexplainable activity with reference to some rate of effective "spontaneous" occurrence (see the discussion in minor section 3.12.6.2). In all the cases I consider, these background processes are uniform over time (and were I to consider space, they would be uniform across space as well). If one can discern reliable non-uniform structure in observations attributed to a background processes, then that is one data pattern that suggests the existence of a hidden cause.

A hidden cause/entity/event/process cannot be directly observed, but is postulated by the theory as a hypothetically observable entity (even if it is impossible to *actually* observe that entity even according to the theory). An example would be the exact time at which an infection by some pathogen occurs, after which contagion to others is possible but which has an unknown incubation period before the symptoms of the caused disease are actually observed (see, section 3.12). Potential hidden causes will need to be well formed and potentially integrable (if not actually integrated) with other theoretical elements. To be theoretically relevant to any phenomenon, it will have to bear on the observed or posited phenomenon. Consequently, the theory can postulate the occurrence of hidden causes to explain observed or hypothetical phenomena or simulate occurrences as part of inference. Given this role, hidden causes may be used as accounts of unseen causal mechanisms within theories.

A generator produces an effect where there otherwise would be no effect. This relies on the possibility of the non-existence of an effect, which means that there is some implicit asymmetry in the event space and a base-case in which a feature is absent. Generators make the most sense when the effect is an binary additive feature. In discrete time for binary cases, if a generator and a background process produce an effect at the same time step, the effect only occurs once (i.e., the potentially observed events do not represent the hypothetically generated events). Thus a hypothesised background cause weakens the potential ability for generators to demonstrate their effects.

A preventer eliminates an effect where there other otherwise would have been an effect. Thus, similar to generators, preventers require an additive features, but they have a different

relation with background processes. There needs to be a generative background process for a preventer to meaningfully exist. That is, an event must at least plausibly occur in order for its non-occurrence be notable or even sensical. This is made yet more salient when we consider this with regards to point processes in continuous-time, where most of the time nothing will occur. For a preventer to demonstrate itself as a preventer of points in continuous-time, the preventer must be present while the event occurs. If we assume that no two events occur at exactly the same time, then preventers will have to exist over intervals that intersect the events generated by other processes (including background processes) in order to have any effect whatsoever.

3.5 Formal background

The formal/computational/mathematical literature on which this work is based is extensive. Below I give a brief overview of some of the main topics and references to point to works with greater detail than can be provided here.

3.5.1 Directed Graphical Models

As discussed in Chapter 1 and in greater detail in Appendix B, directed graphical models are efficient and convenient ways of encoding probabilistic and causal dependencies. They often have a great deal of semantics imbued into their relations. This can include plate semantics, which allows duplicating the nodes within a plate a certain number of times so that the nodes within a duplication are dependent upon each other but not dependent on nodes outside of the plates (except indirectly, such as through parent nodes that govern all of the plates). They also often have convenient computational and algorithmic properties by allowing probability distributions to be factored into a smaller number of parameters than would be required to support all of the nodes otherwise.

3.5.1.1 Bayesian Networks

Bayesian networks are a variety of directed acyclic graphical model that allows expressing joint probability distributions efficiently in terms of conditional probability distributions. They derive their name from their reliance on these conditional probability distributions as understood through "Bayes" theorem which, put most simply, states that,

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_{\forall a \in A} P(B|A=a)P(A=a)}.$$

In this case we want to update our beliefs about A given our knowledge of B. The distribution over A from before updating $P(A)$ is called the prior distribution. The $P(A|B)$ distribution is called the posterior distribution. $P(B|A)$ is the likelihood function describing the probability of observing values for B given values of A. All of the denominators are normalising constants – various ways of representing $P(B)$ – that ensure that $P(B|A)$ is a well-formed probability distribution.

Often in our contexts we will have many layers of this kind of structure that we need to work through. The conditioning can occur on any random variable, and the marginalisation (summing) operation used to remove A from the normalising constant can also occur over any random variable.

In a Bayesian network, nodes represent random variables whose distributions are defined as part of the total stochastic system given by a populated Bayesian network. Edges go from "parents" to their "children" and in aggregate these define the conditional independence relations that hold between different nodes. A Bayesian network satisfies the local Markov property, where every node, conditional on its parents (but not conditional on its descendants) is independent of all non-descendant nodes. Thus, a node that has no parents is marginally independent of all other variables (though if one of its descendants is given, that independence may no longer hold).

One challenge for learning the structure Bayesian networks from statistical patterns is that many graphical structures are isomorphic to one another. Traditionally, identifying the directionality of an arrow relied on the existence of "V"-shaped triples ($X \rightarrow Y \leftarrow Z$) as they are distinguishable from other triple structures (e.g., $X \rightarrow Y \rightarrow Z$ or $X \leftarrow Y \rightarrow Z$).

3.5.1.2 Causal Bayesian Networks

Causal Bayesian networks are Bayesian networks outfitted with an additional "intervention" operator usually called a do operator. An intervention (in this sense, at least) sets a node to a particular value, thereby rendering it independent of its parent nodes. The influence of this intervention may flow down-stream (at least, to those nodes that are not already observed) which can provide useful information about the directionality of different causal arrows. Indeed, in-

terventions can allow whole new information theoretic notions that map the down-stream effects of an intervention (for more on this see the discussion of causal information flow in minor section 1.2.4.2, Ay and Polani [29], and Nielsen et al. [19]).

3.5.2 Causal theories as in Griffiths and Tenenbaum [1]

Griffiths and Tenenbaum [1] lay out an account of theory-based causal induction that greatly influences the rest of the work reported here. They are motivated by similar concerns: human cognition In it they provide a mechanism for defining nonparametric probability spaces for causal graphical models (including causal Bayesian networks). That is their *causal theories* generate a hypothesis space and probability measures over the elements of that space, which may be unbounded, by specifying ontologies, plausible relations, and functional forms. The ontology states which types of things exist in the world, what properties they have and what kinds of relations (predicates) can exist between different types of things. Plausible relations state the probability of different kinds of relations actually occurring, and functional forms describe the form of that relation in terms of mathematical or logical constructs and the probability that any of those values occurs. In this they are able to describe traditional causal inference problems with contingency table type data, but are able to do so for a variety of different tasks where different kinds of models would need to be invoked (including varieties of deterministic models as well as the generative and preventative models that I will discuss here). They also are able to handle relations that exist over spatial and temporal dimensions as well as domain specific and cross domain causal reasoning.

Though this chapter is rooted in this paradigm, it goes far beyond what was originally discussed in terms of specifying the notions necessary for a complete theory of continuous time causal induction. The goal of causal theories is to encompass a wide scope of phenomena with a general framework. The work here points out that causal induction with temporal data, in particular causal induction with continuous-time data, requires a great deal of further specification and care than could be given in that work. Additionally many of the problems that I address have other features that were not covered in the original work (e.g., generating large numbers of graphs and then filtering them on the basis of post-hoc graph theoretic calculations, developing a formalism for sampling the entire history of a finitary Poisson process, and addressing the distinction between events that occur many times from many processes and one shot processes). Because of this emphasis on the problems that arise specifically in the context of *continuous-time* causal induction and because of its roots in the *causal theory* framework

described in Griffiths and Tenenbaum [1], I have dubbed the framework *continuous-time causal theories*, or CTCT[✧] for short.

3.5.3 Probabilistic Functional Forms: noisy-or & noisy-and-not

Though, in theory, Bayesian networks can express arbitrary relations between arbitrarily structured variables, in practice nodes tend to express particular classes of relations between particular classes of variables. Identifying these functional forms I will focus on the role of binary variables as they most directly relate to the analogous studies on contingency tables (for a more general introduction, see Murphy [174]). In fact, I will focus not only on binary variables but binary additive (as in not substitutive) variables, as the distinction of presence and absence are key to defining the semantics of these commonly used relationships[✺]. When the cause and effect are binary additive variables two functional forms are commonly used that are noisy versions traditional logical relations: noisy-OR and noisy-AND-NOT.

Suppose you had a system with one background (generative) cause[*] and one other cause with one common effect. If that other is a generative cause, then we can define both causes in terms of the probability that they will individually generate the effect. Of course, there is the possibility that on any particular trial they would both generate the effect[✾], and that must be taken into account. Rather than constructing the probability that they will produce the effect in aggregate, it is easier to calculate the probability that it is *not* the case that both of them *failed* to produce the effect (and therefore demonstrating that at least one of them did). That is, if the background cause B has the probability w_b of producing the effect, and the generative cause G has probability w_g of producing the effect when the cause is present, then the probability of the

[✧] I considered for a while of using CT², but it felt like it would at worst be confusing, and at best difficult to convey reliably across filetypes and formats.

[✺] One can have substitutive versions of functional forms that expect presence and absence, but then some assignment that maps variables from the substitutive domain such as "left" vs. "right" and "absence" and presence. The meaning of the relation changes dramatically depending on how you accomplish that mapping, making substitutive binary variables equivalent to many different causal relations. In that case, one would often use a generic model (in the vein of Lu et al. [78], Griffiths and Tenenbaum [1]) that gives a separate probability for each combination of parent values.

[*] It is possible, but somewhat more complicated to have a background preventative cause. One in that case needs to assume that there is some potentially present non-background generative in order for that background preventative cause to be observable. Given that that seems to run against what we normally mean by "background", I will not explore the topic further here.

[✾] The absence of this possibility is one of the key distinctions between continuous and discrete time causal systems (at least in the generative case).

effect when the cause is present is $p(E|G=1,B=1) = 1-(1-w_b)(1-w_g)$ and note if the cause is not present $p(E|G=0,B=1) = 1-(1-w_b) = w_b$. Note that B is always 1 because it is the background and is therefore always present. Together, these can be written as $p(E|G=g,B=1) = 1-(1-w_b)(1-w_g)^g$. More generally for one event and k generative causes (and one background cause with probability w_b) a Noisy-OR causal system looks like:

$$p(e|\{G_i=g_i\}_1^k, b=1) = 1-(1-w_b)\prod_{i=1}^{k}(1-w_{g_i})^{g_i}.$$

The situation for noisy-AND-NOT is similar. We still need to assume a generative background cause with probability of producing the effect w_b. This time, though, what we want to know is the probability that, when the background cause generates the effect, the preventer also occurs and thereby prevents the effect. It does not matter what the preventer does if there is no event for it to cancel[Δ]. So, if the background cause has the probability w_b of producing the effect, and the cause has probability w_p of canceling the effect when the cause is present, then the probability of the effect when the preventative cause is present is $p(E|P=1,B=1) = w_b(1-w_p)$ (or the joint probability that the background cause occurs and the preventer does not occur). And, again note if the cause is not present $p(E|P=0,B=1) = w_b$, making value agnostic functional form $p(E|P=p,B=1) = w_b(1-w_p)^p$. The case with m preventative causes can be similarly defined as the joint probability that the background cause occurred and that every one of the preventers did not occur:

$$p(E|\{P_i=p_i\}_1^m, B=1) = w_b\prod_{i=1}^{m}(1-w_{p_i})^{p_i}.$$

These can be combined into a joint generative-preventative form by considering that the w_b is merely a probability that the effect occurred, and can be replaced by any equivalent probability function that determines how likely the event is to occur from that part of the function). In mathematical terms this is:

$$p(e|\{G_i=g_i\}_1^k, \{P_i=p_i\}_1^m, b=1) = \left(1-(1-w_b)\prod_{i=1}^{k}(1-w_{g_i})^{g_i}\right)\prod_{i=1}^{m}(1-w_{p_i})^{p_i}.$$

[Δ] It is in this sense that the presence-absence distinction in binary additive variables is crucial for using these functional forms properly. Without it, this asymmetry between the values of 0 and 1 make little sense.

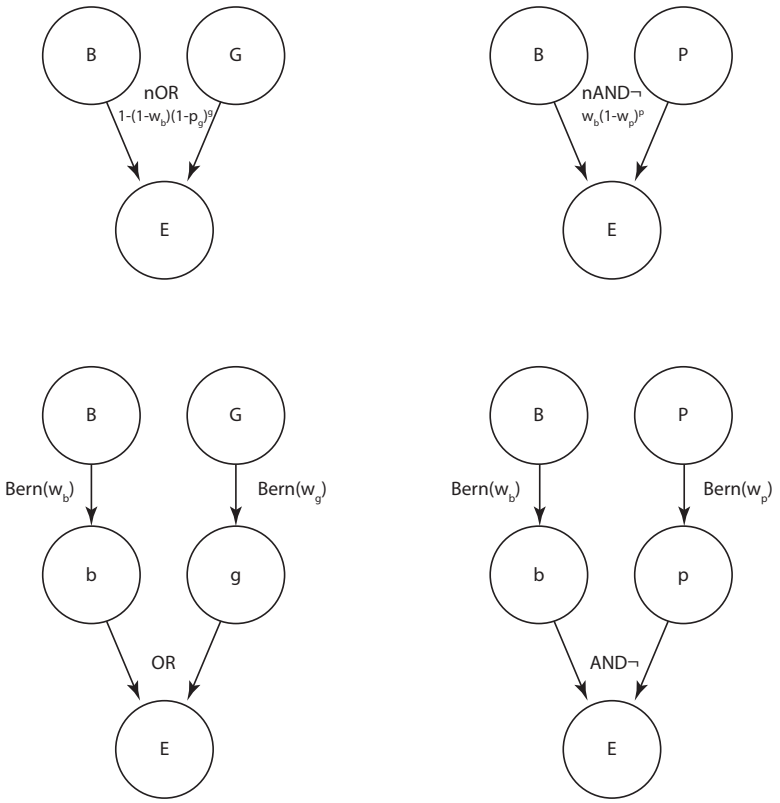


Figure 3.3: The noisy-OR and noisy-AND-NOT functions can be interpreted directly in terms of a stochastic-logical algebra, or in terms of a noisy gate on presence of the cause that multiplies it by a Bernoulli random variable and pushes that output into a logical algebra.

Noisy-OR and noisy-AND-NOT can be seen as probabilistic versions of a logical OR and AND-NOT (see Figure 3.3). This handles the case for a single trial, and even for a countably infinite set of trials (i.e., a Bernoulli process). However, we are attempting to move away from countable identical trials and toward uncountable infinite time. What will be useful is if we can extend these notions that are defined in discrete time to formal structures that are well defined over continuous-time.

3.5.4 Stochastic Processes

A stochastic process is a way of defining a probability distribution for a collection of random variables[175]. This includes defining probability distributions for infinite collections of random variables[¤]. This holds for any manner of defining the distribution for an infinite set of variables (e.g., indexing them over the 2-d Euclidean plane), though they often are considered in terms of an time index $\mathcal{T}$.

Stochastic processes give a well-formed way to define statements such that every random variable derived from this process will be independent and identically distributed (i.i.d.). Thus you can see many standard probability and statistical problems in terms of statements about stochastic processes. This is useful because it gives a rigorous sense in which we can define generalisation from a set of (presumed) i.i.d. samples from some process to describe that process going forward (assuming that the i.i.d. property still holds). That said, the most interesting properties of stochastic processes arise when one attempts to build dependencies between the random variables across indices.

A discrete-time stochastic process defines a probability distribution over a countably infinite set of variables. This is the domain in which most trial based analyses are implicitly defined over. People assume that there will be a series of well defined bounded events and attempt to calculate statistics over those events as if they are independent across time, once all of the relevant features at a particular time step are taken into account. Bernoulli processes are one such discrete-time stochastic process. Markov chains are an interesting discrete-time process that see wide use in modern probabilistic modelling[174].

A continuous-time stochastic process defines a probability distribution over an uncountably infinite number of moments. This includes Gaussian processes and continuous time Markov processes (also known as continuous time Markov chains). What I will concern myself with are Poisson processes.

3.5.5 Poisson Processes

Poisson processes are stochastic point processes. That is, they describe the random occurrence of infinitesimal points over some underlying support space, such as the positive real line R^+ or a 2-d Cartesian plane (R^2). For an excellent introduction to the general theory of Poisson processes, I recommend Kingman's[176] book *Poisson Processes*.

[¤] Throughout this I will be implicitly referring to Ross[175], which I recommend as an excellent textbook on stochastic processes.

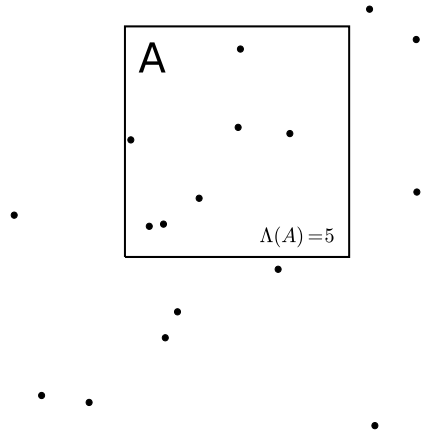


Figure 3.4: Realisation of a Homogeneous Poisson point process over a 2-d Euclidean space. The expected number of events to occur within the bounding box labelled A is 5.

Poisson processes can be interpreted in a number of ways that lend themselves more or less easily to different applications.

In spatial domains[¢], it is easiest to think of the Poisson process as a stochastic point process which has an underlying intensity measure that tells you how many points are expected to occur over any subspace. That is, there are infinitesimal points randomly distributed over a space, where for a subspace A you expect $k_A \sim \text{Poisson}(\Lambda(A))$, as in Figure 3.4, where $\Lambda(A) = 5$.

However, I will be focusing on a temporal version of the Poisson point process which has some additional features due to all points existing on a single dimension. Most of these properties are included in the counting process picture of Poisson processes, that emphasise the (usually) total ordering present when looking at events that occur along a single dimension[$].

To describe these systems, we need a notion of a intensity function $\lambda()$ and an intensity measure $\Lambda()$, where the intensity function takes points as arguments and the intensity measure takes intervals of times as arguments. The intensity measure is just the integrated intensity function over the interval in question $\Lambda([t_1, t_2)) = \int_{t_1}^{t_2} \lambda(s)ds$.

[¢] I introduce Poisson processes in terms of a spatial domain on the recommendation of Kingman [176] who points out that the particular features of the real line(R^+), specifically that the points can be ordered, obscure some of the generality and the simplicity of the process.

[$] The ordering is not total if one is studying Markov jump processes or non-simple/non-orderly processes, which have more than one event that is counted/occurs at the same time.

In the temporal case, the notation often used for describing the general likelihood for the number of arrivals ($N()$) that occur during the period of time $[t_1, t_2)$ is:

$$\mathcal{L}(N([t_1, t_2)) = k) = \frac{\Lambda([t_1, t_2))^k}{k!} \exp(-\Lambda([t_1, t_2])), \tag{3.1}$$

and in the homogeneous case, where the rate everywhere is constant λ_0, has a likelihood of the form:

$$\mathcal{L}(N([t_1, t_2) = k)) = \frac{[\lambda_0(t_2 - t_1)]^k}{k!} \exp(-\lambda_0(t_2 - t_1)). \tag{3.2}$$

I will describe Poisson processes in two senses: the arrival process sense and the rate-of-events sense[€]. Sequences of point events can be described sequences of successive arrivals, making the arrival perspective convenient for computing point event likelihoods. But it is easiest to conceive of the causal effects of point events in terms of their altering event rates. Both perspectives prove useful.

Arrivals. The arrival sense can be understood by anyone who has ever waited in a queue. You will have to wait some amount of time before your turn, and we can assign a probability that you will be served by time t. If you were next in the queue and it were governed by a homogeneous Poisson Process with rate λ, the waiting time distribution of being served by time t would be an exponential distribution with mean $\frac{1}{\lambda}$ ($t \sim \text{Exp}(\frac{1}{\lambda}) : p(t) = \lambda e^{-\lambda t}$). Interestingly, this distribution is *memoryless*, such that, regardless of how long we have waited, we still expect to wait the exact same amount of time — it has no memory of how long it has been since the last event. This memorylessness property does not hold for the general class of nhpps.

Rates A sequence of events that arrive according to a sequence of waiting times can be seen to be equivalent to the event-rate perspective of the Poisson process. If you have events embedded in some space, and take the number of events that were expected to occur in some subspace, that can be thought of as a (noisy) measure of the rate of events over that subspace. Equivalently, we can count the number of events that occur in a measurable time-period, rather than looking at the delays between each event. Poisson processes define a "rate" of

[€] Poisson processes can be defined over higher dimensional spaces (e.g., R^3) than the real line. This complicates the arrival perspective, which implicitly relies on the order that events "arrive". The event-rate perspective is unchanged in higher dimensions; in that sense it could be said to be more "fundamental" than the arrival perspective. In this chapter, I focus on processes defined over time ($[0, \infty)$).

events, which describes the expected count of events to occur in any interval. A homogeneous Poisson process has a constant rate, λ, and for a time interval with length $|\tau|$ we can expect to see event-count distribution governed by a Poisson random variable, with mean $(\lambda|\tau|)$. In the case of nonhomogeneous Poisson processes, we will have a rate-*function* defined over time $\lambda(t)$. Integrating this function over some time-interval defines how many events are expected to occur (i.e., for $\tau_{[a,b)}$ the expected event count is $\int_{\tau_{[a,b)}} \lambda(s)ds$).

Arrivals and Rates The arrival perspective provides a probability distribution over intervals of time (i.e., intervals defined from now until the next event arrives), while the rate-of-events perspective provides an instantaneous measure of event likelihood which is comprehensible only in terms its integration over intervals of time. The former is more useful in cases where events are analysed one at a time. For example, when simulating dependent event sequences or calculating the probabilities of event sequences in terms of the likelihood of each event's occurrence given the previous relevant occurrences. In my model of Lagnado and Speekenbrink[7], I will use this perspective to define the likelihood of each inter-arrival period conditional on the previous events.

The rate-of-events perspective is useful when simulating many events when the rate is independent of the particular occurrence of the events. The rate perspective is also useful for calculating event likelihoods, when the interval during when the events occurred is known, but the exact occurrence times are unknown. Pacer and Griffiths[156] use this technique to analyse the data given to participants in Greville and Buehner[2] in which data were presented in a tabular form that described the day during which bacteria died but not the exact timing of the events. This property allows recovering a trial structure from continuous-time by integrating over intervals of time and treating occurrences within those intervals as events that occurred in those trials.

Most importantly for our uses, it is most straightforward to see causes as altering the rate-of-events and then computing an expected wait-time distribution based on those altered event rates. Describing effects in terms of rate changes will be the key to the *causal* aspect of my framework. Fortunately, Poisson processes have two closure properties, superposition and thinning, that allow creating continuous time analogues of noisy-or and noisy-and-not.

3.5.6 Non-homogeneous Poisson Processes

Not all Poisson processes have the same rate at all time (or over all spaces). Processes that have rates that vary over time (or space) are called non-homogeneous Poisson processesNHPP.

3.5.6.1 Piecewise homogeneous Poisson processes and time dilation

The simplest case of a NHPP is where it is a piecewise homogeneous Poisson process. Over some time intervals it will have different rates than others, but within each piecewise defined part of the rate function it will be constant. Thus as long as you analyse time intervals that stay within those piecewise constant parts, the process will be governed by the general rules for homogeneous Poisson processes.

This leads to one of the simplest interpretations of a NHPP in terms of time dilation. Suppose you had two homogeneous processes(PP_1 and PP_2) where the rate of one process was twice that of the other one ($\frac{\lambda_1}{2} = \lambda_2$). You can treat the faster process PP_2 as having had twice the time over which to sample for each equivalent unit of time had by the slower process PP_1. This means that if you had a sampling scheme for generating events according to PP_1, you could use that same sampling scheme for PP_2 for the same amount of time by allowing it to run for twice the time, and then scaling the times of occurrence by $\frac{1}{2}$ to ensure that the time-scales are appropriate. This is how the rate-inversion method of sampling NHPPs works: you sample from a homogeneous Poisson process with a rate of $\lambda = 1$ and then invert the rate function defining the NHPP and apply that function to determine when the equivalent events occurred.

As you might guess, this time-dilation approach to sampling from NHPPs can be extended to computing other features of the process. It also can be used in NHPPs that are not piecewise constant as long as their intensity measure $\Lambda()$ is invertible[176]. If our intensity measures are finite over a finite part of the support space, and our intensity functions are càd-làg ("continue à droite, limite à gauche", or "continuous on the right, limit on the left") – that is, they will always have a limit when approached from the left, and will always have a limit that is equal to the value of the function when approached on the right (this is why we use left closed, right open intervals $[x, y)$ in defining our time intervals) – then we will have a monotone transform (a transform that does not change the order of the points, merely their metric distances from each other) that will convert a NHPP to a homogeneous poisson process.

[] More precisely, it is not whether $\Lambda()$ is invertible, but whether the function $M(t) = \Lambda(0, t) = \int_0^t \lambda(s)ds$ is invertible in the description by Kingman [176]

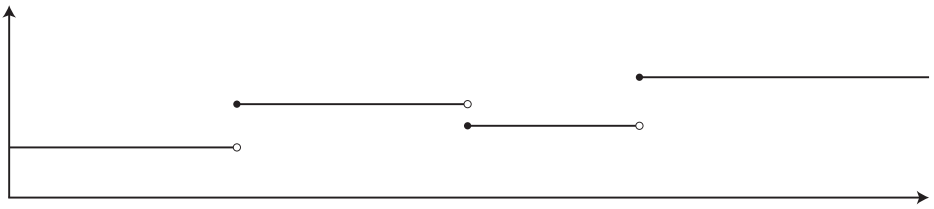


Figure 3.5: A càdlàg function has limits when approached from the left and is continuous when approached from the right (it has a limit and that limit is equal to the value of the function at that point).

3.5.6.2 Likelihoods in the case of NHPPs

The likelihood for a non-homogeneous Poisson process has the same form as that for a homogeneous Poisson process, that is it is the same as Equation 3.3. Unlike the homogeneous case, we cannot simplify the general formula into an expression in the nonhomogeneous case; there is no analogue to Equation 3.2 that holds for NHPPs in general.

However, if we have a càdlàg intensity function, we can simplify the expression to some degree, at least for computational purposes. These will arise whenever there are jumps in the intensity at a particular moment in time. This will prove useful for introducing new processes that are then combined with a base process using operations like superposition and thinning (as described in subsection 3.5.7).

If the rate function $\lambda(\cdot)$ has m jumps at $\{t_1, t_2, \dots, t_m\}$ then,

$$\begin{aligned}\mathcal{L}(N([t_1, t_m)) = k) &= \frac{\Lambda\left([t_1, t_m)\right)^k}{k!} \exp(-\Lambda([t_1, t_m])), \\ &= \frac{\left[\sum_{i=1}^{m-1} \Lambda\left([t_i, t_{i+1})\right)\right]^k}{k!} \exp\left(-\sum_{i=1}^{m-1} \Lambda([t_i, t_{i+1}])\right). \qquad (3.3)\end{aligned}$$

One way one can produce NHPPs almost accidentally is by trying to reason about systems of Poisson processes that do not initialise at the same times. If one disallows more than one event to occur at a particular point in time, and one sees these processes as induced causal effects then NHPPs can be expected to arise in causal processes involving even only homogeneous Poisson processes.

3.5.7 Superposition and Thinning in Poisson Processes

This section will develop a rough physical model to aid in thinking about NHPPs as formed by functions on homogeneous Poisson processes. Namely, I aim to provide an intuition for the superposition and thinning closure-properties of homogeneous Poisson processes from the rate-of-events perspective. I do so by sketching a mechanistic picture of a particle emission system that exhibits these properties.

First, consider a decaying radioactive material which releases particles at a constant rate, ψ. With a particle detector around the material you record the time-stamp at which particles hit the detector. A particle is expected to hit the detector, on average, every $\frac{1}{\psi}$s. This detector will then be recording a homogeneous Poisson process with rate ψ.

Suppose you were to place a barrier to block some of the paths leading from materials to the detector (call the proportion blocked $\nexists : 0 \leq \nexists \leq 1$, as in the parameter associated with the orange filter in Figure 3.6). From the detector's perspective, events associated with particles blocked by a filter are events that never occurred. This process is known as filtering the Poisson process, and if $\nexists$ is independent of the generating process, filtering gives a Poisson process with rate $(1 - \nexists)\psi$.

Suppose you were to place another radioactive material in the detector, of a different kind than the original, but which did not interact with the original radioactive material (see the blue and green materials in Figure 3.6). From the perspective of the detector, there is no difference between the particles hitting it from different materials — it only knows *that* a particle hits it and *when* it hits it. If we suppose the rate of this new material's emitting particles is constant at ψ_1, then the total set of events would be a Poisson process with rate $\psi + \psi_1$. This is the superposition property of Poisson processes: if jointly independent, the union of the events from two Poisson processes will be a Poisson process whose rate is their sum.

Superposition and thinning allows us to see how the rates of Poisson processes can change without altering their underlying structure. We can apply these transformations at particular times or intervals of time, thereby producing and increase (via superposition) or a decrease (via thinning) in the rate of events while at all times maintaining its identity as a Poisson process. Applying superposition or thinning as time-dependent functions thus allows one way to create NHPPs that nonetheless can be understood in terms of component processes and their transformations.

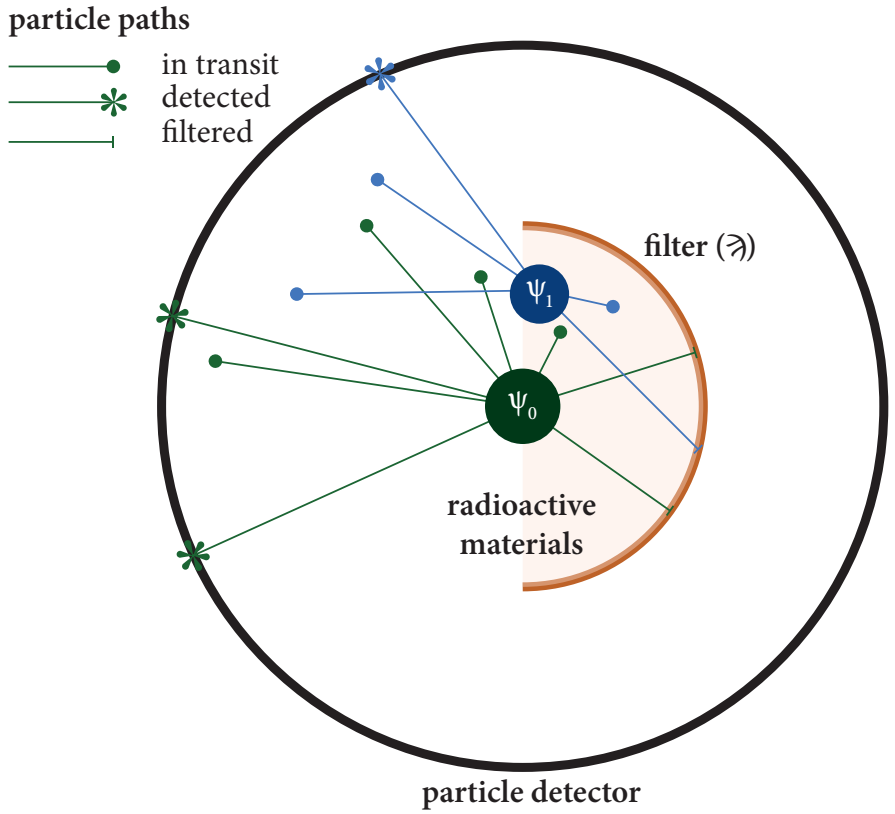


Figure 3.6: Particle emission detector model for visualizing superpositioning and filtering Poisson processes. Colour distinguishes particle origins prior to detection, with colour lost in detection. Filtered events are never detected.

3.5.8 Properties of Poisson processes

There are some useful properties of Poisson processes, some that hold in general and some that are convenient and I will assume hold for the purposes of my work.

3.5.8.1 Defining first arrivals as minima

When we consider Poisson processes from the wait-time perspective, we will be generally only concerned with the amount of time that passes until the first event associated with the process occurs. We know that for a homogeneous Poisson process the wait-time is a exponential distribution. Exponential distributions are convenient for these purposes because the minimum of an arbitrary set of exponential distributions is also exponentially distributed. We can use this

fact to determine the wait-time until the first event from an arbitrary number of homogeneous Poisson processes.

It is useful to note that the first event will be defined by the minimum of the exponentially distributed wait-times. Specifically, if we have n exponentially distributed random variables $\{X_1, X_2, \dots, X_n\}$ with rate parameters equal to $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$, we want to find the distribution of $X_{\min} = \min(\{X_1, X_2, \dots, X_n\})$. Normally we would approach this by computing the cumulative distribution function for the minimum $F_{X_{\min}}(x)$. However, as is common in the calculation of order statistics, using the complementary cumulative distribution function $1 - F_{X_{\min}}(x) = \bar{F}_{X_{\min}}(x) = P(X_1, X_2, \dots, X_n > x)$ and solving backward is more useful, because the minimum value will be less than all of the other values (by definition) and thus equal to the value that all of the random variables are greater than[⁛] With this and $F_{X_i}(x) = 1 - \exp(-\lambda_i x)$ in hand we can compute,

$$\begin{aligned}
P(X_1, X_2, \dots, X_n > x) &= \prod_{i=1}^{n} P(X_i > x), \\
&= \prod_{i=1}^{n} 1 - P(X_i \leq x), \\
&= \prod_{i=1}^{n} 1 - (1 - e^{-\lambda_i x}), \\
&= \prod_{i=1}^{n} e^{-\lambda_i x}, \\
&= e^{-x \sum_{i=1}^{n} \lambda_i},
\end{aligned}$$

which is an exponential random variable with rate equal to $\sum_{i=1}^{n} \lambda_i$. Thus, the expected amount of time until the next event would be $\frac{1}{\sum_{i=1}^{n} \lambda_i}$.

How do we handle the superposed Poisson processes (especially in the case where it is not even piecewise homogeneous)? One approach relies the independence properties inherent to Poisson processes. These in general make the distribution fairly easy to calculate for even this first event[⁛].

[⁛] You can say strictly greater than because in a continuous domain the probability of a random variable taking any particular value is (usually) 0 and does not need to be taken into account.

[⁛] Ostensibly they should also make it easier for the other order statistics based on the order statistics of independent exponential random variables[177]. However, I have found applying these results on the order statistics of independent and not identically distributed (inid) exponential random variables

If we consider NHPPs from the time-dilation perspective, then we can look at converting each of the NHPPs into a common base unit and then generate the first event from each of those processes and take the minimum of those generated. A slight caveat is needed if the NHPPs have a finite measure over their full support space, as then they could produce an sample with no events (in which case there is no first event). In fact, this case will be useful for different reasons as will be discussed in subsection 3.7.6.

3.5.8.2 Assigning identity to events from superposed processes: an application of total thinning

Once we know an event occurs due to some combination of processes, we can identify the probability that the event arrived from either by taking the ratio of the intensities of the processes at the time of occurrence. I.e., for three processes A, B, C that are superposed to contribute to the total process X the probability that the arrival t on X is from A is

$$p(\{t\} \leftarrow \mathbb{t}_A) = \frac{\lambda_A(t)}{\lambda_A(t) + \lambda_B(t) + \lambda_C(t)}.$$

And similarly the probability of coming from B is $\frac{\lambda_B(t)}{\lambda_A(t)+\lambda_B(t)+\lambda_C(t)}$ and C is $\frac{\lambda_C(t)}{\lambda_A(t)+\lambda_B(t)+\lambda_C(t)}$.

The key thing to notice is that this relies only on the instantaneous intensities $\lambda()$ not on the intensity measures ($\Lambda()$). If one were to know that an event occurred over an interval of time without knowing the exact time, the ratio would then be between the intensities:

$$p(t \in [t_1, t_2), t \leftarrow \mathbb{t}_A) = \frac{\Lambda_A([t_1, t_2))}{\Lambda_A([t_1, t_2)) + \Lambda_B([t_1, t_2)) + \Lambda_C([t_1, t_2))}.$$

One way to interpret the validity of this action is to see how the transformation can be reversed through thinning. Consider first the total process from the superpositon of m component processes with rate ψ_{total} (i.e., $\psi_{\text{total}}(t) = \sum_{i=1}^{m} \psi_i(t)$). Then thin that total process into m separate processes such that the rate of thinning for each of the m components at all times is equal to the ratio of that process' contribution to the total rate against the total rate ($\eth_i(t) = \frac{\psi_i(t)}{\sum_{j=1}^{m} \psi_j(t)} = \frac{\psi_i(t)}{\psi_{\text{total}}(t)}$). Because a an independent thinning process produces (in expectation) another Poisson process with rate equal to the product of the thinning value and the total rate the resulting processes would have rates $\psi_i'(t) = \eth_i(t)\psi_{\text{total}}(t)$. By substituting in

challenging to interpret when the exponentials do not have the same "starting point". Furthermore, it is unclear whether you could use that form to mathematically derive a closed form expression for an arbitrary set of superposed NHPPs. I have not yet figured out how to state it generally. Fortunately for the forward sampling algorithm, you only need the identity of many first events

$\frac{\psi_i(t)}{\psi_{\text{total}}(t)}$ for $\lambda_i(t)$, you can see that (after cancelling the denominator of the fraction by the total rate) that $\psi_i'(t) = \psi_i(t)$.

3.5.8.3 Simple and orderly point processes

A simple point process is a process in which at most one event is expected to occur at any moment.

$$P(N(\{t\}) \in \{0, 1\} \forall t) = 1 \tag{3.4}$$

Because infinitesimal moments are considered problematic in analysis, a more rigorous statement is that of orderliness which is defined in terms of "little-*o*" notation. "Little-*o*" notation states that one process($f(x)$) grows more slowly than another ($g(x)$), to the point where in some limit(usually as $x \to \infty$) the first process $f(x)$ becomes negligible to the second process. For the case of point processes, we want to ensure that there is never more than one event at a particular time point. In order to speak not of moments, but of intervals we want to state that as an interval gets arbitrarily small, the probability that more than one event occurs in that interval is much smaller than that interval. In more formal terms:

$$P(N((0, h] \leq 2) = o(h)(h \downarrow 0). \tag{3.5}$$

That is, as the measure of the interval of time approaches 0, the probability that two or more events will occur in that interval approaches 0 even more rapidly than the measure of the interval.

In general the processes I am going to consider are going to be of the simple or orderly varieties. One of the reasons that I do this is because of its consequences for causal interpretations of continuous-time processes — specifically, by only assessing these processes we can be assured that all processes will unfurl into a directed diagram. That is, loops are even less of an issue than they are for discrete time versions of temporal causal graphs. They fall directly out of our primitives rather than needing to be explicitly enforced. Inconsistent loops simply cannot exist.

3.5.8.4 *Dirac δs a pointedly useful function*

The Dirac delta function[M] was introduced by Paul Dirac [178] to be able to represent point masses in quantum mechanical theory. The two-argument $\delta(\cdot, \cdot)$ has an infinite spike where the two arguments agree and 0 elsewhere. This can be seen as a version of the one-argument Dirac delta with modified notation, where the input to $\delta(\cdot)$ is $\delta(x - y)$ for $\delta(x, y)$.

$$\delta(x) = \begin{cases} +\infty & x = 0 \\ 0 & x \neq 0 \end{cases} \tag{3.6}$$

One reason to prefer the Dirac delta is that it is easier to express the notion that $\int_{-\infty}^{+\infty} \delta(x)dx = 1$ without needing to privilege either of the two arguments as being the one that will be integrated over.

3.6 Desiderata for continuous-time causal theory based framework for causal induction

There is great variety in the phenomena available to a framework for continuous-time causal induction. In subsection 3.14.2, I describe an even larger class of these features that could potentially be fulfilled

Thus, it will be helpful to identify the most vital features for allowing the framework to capture a wide class of these cases. The following sections detail an important set of these properties.

3.6.1 Ontology

We will need an ontology that provides the tools needed to express the range of processes possible in a continuous space. This includes both points and intervals and the relations needed to relate them in a coherent way.

[M] Technically this is not a function, because it integrates to 1 but is 0 almost everywhere. True functions that are 0 almost everywhere need to integrate to 0 when a Lebesgue integral is applied to them

3.6.2 Plausible continuous-time *sets* of relationships

In the original causal theories paper, the plausible relationships needed only to be defined singly; that is, you could provide a statement about whether a relationship existed between any two entity nodes (including hidden entities) regardless of the large scale consequences of the assignment of relationships. This is not enough for handling continuous time models.

3.6.3 Generative and preventative causal relations.

It is vitally important when modelling human causal inference to distinguish between causes that generate effects and causes that prevent effects[30,1]. People make dramatically different predictions based on which type of relationship they are looking for. Thus, we would want the framework to be capable of doing the same. In discrete time, Griffiths and Tenenbaum[30] used the Noisy-or and Noisy-and-not logic gates to represent a cause that generates or prevents effects with reference to a background rate of the effects' occurrence. Because these discrete time parameterisations will not hold in continuous time, we will have to redefine what we mean by a generative and a preventative relation for continuous time.

3.6.4 Persistent, decaying effects

In most models of causation that work in discrete time or over trials in which events are treated as occurring simultaneously, a cause can only influence an effect if and only if that cause is present on a particular trial. This is undesirable if we are to develop a framework for continuous-time causal inference, particularly if we are to allow point causes. These causes occur only instantaneously, which would make their causal influence infinitesimal if simultaneity were required. Thus we need to have some way to extend an event's influence over the time following its occurrence, ideally using the same mathematical primitives that we use to describe interval causes.

3.6.5 Intervention.

The framework should capable of considering interventions in the sense meant in causal graphical models with simultaneity assumptions[28] That is, in that case, an intervened upon node is said to be rendered independent of its parent nodes and takes on whatever value with probability 1. Work has extended the notion of intervention to dynamic Bayesian networks (or causal graphical models for time series, though only discrete time series were considered) and to cases

beyond atomic intervention including random interventions that set a value stochastically according to some distribution[⊛ 179].

3.6.6 Multiple independent causes and multiple effects

In the case where there are many causes, if we assume they are (at least conditionally) independent, their joint effect should be able to be understood in addition to their marginal effects. In the standard causal Bayesian network (discrete process, with binary variables), without assuming possible functional forms, this requires 2^k separate parameters for k parents with one child. In my work, I want to avoid this explosion of parameters. This is one reason for defining functional forms in terms of generators and preventers that can be composed with existing distributions.

Similarly, especially in the case of potential feedback loops, we will need to have a way to define effect relationships differently along different directed edges. These edges should be able to have different temporal properties. We may even have cases where we need to model more than one arrow from node X to node Y (one that is generative and one that is preventative). We will want to have a way to specify how different kinds of connections will be managed in relation to one another.

3.6.7 Composable, computationally well-formed likelihoods for many kinds of data and relations

Given the known difficulties in dealing with complex stochastic processes like those defined over time[180], it would be good to have straightforward ways of computing the relevant likelihoods for a variety of data types that we wish to model. Ideally these models would be defined such that they could be easily swapped with one another (i.e., that they are composed of the roughly the same kinds of parts).

We want to be able to handle all of the data that could accrue because of our ontology.[◈].

⊛ This essentially sets the distribution of the randomly intervened on node to be whatever the distribution of the intervention is (rather than a point mass on a single value). Eichler and Didelez [179] also cover conditional intervention as an intervention only occurs conditional on the occurrence of other events but on the basis of a decision rule rather than as part of the causal system.

◈ In fact in the most general case, we fail to meet this criterion. However it is unsurprising when you consider the great variety of data and relation types that are feasible within the most general version of CTCTs as described in subsection 3.14.2.

We also want to be able to handle a wide variety of ways of relating largely the same kinds of data. For example, we want our processes to account for one shot processes (those processes whose events can occur only once), as well as multiply instantiated one-shot processes over individual entities. That is, any one process that is instantiated on the entity can only have at most one event, but there can be multiple instances of causes that each induce a new process of that type on the entity.

We want to be able to account for one-to-one cause effect mappings (as is expected by many trial structures and experimental procedures).

We want to be able to represent relations between causal processes that have many events on both the cause and the effect side of the process.

3.7 A framework for continuous-time causal induction

Below are discussions and derivations that build up the pieces of the continuous time causal theory (CTCT) framework. It takes to the point of being able to produce all of the models that I use to account for empirical phenomena, as well as a wide swathe of other considerations.

3.7.1 Ontology: points and intervals

In terms of the events in question, we will need to be able to express both point and interval events, and will focus on processes where the points are the causal effects. In subsection 3.14.2, I discuss the more general problem of defining an ontology for all the events expressible in a continuous-time process. Because we will want to discuss points in terms of rates, we will effectively need a calculus for using point and interval causes to modify or otherwise define real valued functions, ideally càdlàg functions. There are a number of distinctions about how one can interpret interval events that need to be made explicit (as otherwise our application of causal intervals would be fundamentally ambiguous).

3.7.2 Plausible relation sets: graph priors, supergraphs, and post-hoc filters

In order to accommodate the constraints of various continuous time causal systems and the prior knowledge that you might have on it, we need to consider probability distributions being defined over sets of edges, possibly even over the graphs themselves. This is in contrast to traditional causal theories in which the total graph prior for the structure was built from the aggregated probabilities of the individual causal relations.

We need this mechanism in order to account for knowledge you have about which in which you will need to have a probability distribution not just over individual edges, but over sets of edges and even over the functional forms and base-rates on those edges. For an example of where this occurs see section 3.12. In my case, under one set modelling assumptions, we need to be able to respect particular graph theoretic conditions (such as the existence of a path between two nodes) in order to have a valid system over which to calculate. If we do not have that, we will be required to have base-rates for our nodes.

One way to accommodate this is to use a supergraph structure as described in minor section 3.12.6.1.

3.7.2.1 Multi-relational structure of interest: feedback loops

One particular kind of large scale structural feature that we certainly want to be able to employ is feedback loops. Having a way to reason about feedback loops these is an important part of the motivation for using continuous-time stochastic processes in general, and simple Poisson point processes in particular.

By extending these graph theoretic filters to consider the semantics of the relations, this allows ensuring that any feedback loops do not produce patterns of activity that would violate modelling conditions. Also, because we have priors overs sets of edges, if we so desire, we could directly model the probability of feedback loops.

That said, if we wanted to abolish the possibility of feedback loops creating priors over sets of relations that give any graph with feedback loops 0 probability is one way to accomplish it. This means that if we want to return to the case of synchronous state space modelling we can directly enforce a directed acyclic graph condition merely by specifying a certain relation-set prior.

3.7.3 Generative and preventative causal relations: superposition and thinning

The properties of the Poisson process – specifically invariance of the form of the stochastic process under the superposition and thinning transformations – can be used to characterize generative and preventative causal relations. Suppose that there are i generative causes($\{C_i\}$) and j

It may be possible to bake impossibility conditions like these into the likelihood, however it would be superior if we could have a mechanism for adhering to these constraints at the structural level rather than the semantic level.

preventative causes ($\{C_j\}$), and they exist over intervals of time. That is, $\forall C_a \in \{C_i\} \cup \{C_j\}, \exists T_a(C_a = 1) \subset \mathcal{T}$ where $\mathcal{T}$ is the set of all non-measure-zero time intervals and $T_a(C_a = 1)$ is the set of intervals during which C_a occurs. Let the Poisson process PP_0 be a background rate of effect occurrence with an unknown time-invariant rate function $\lambda_0 > 0$. Causes assert their influence by altering the base-rate of the effect.

Generative causes will superpose themselves onto the background process, thereby increasing the rate of effect occurrence. That is, we can think of a generative cause C_i as producing a series of effects on its own, thereby inducing a Poisson process PP_i with parameter $\psi_i(t)$, where we assume that the cause only exhibits a non-zero effect when it is present (i.e., $t \in T_i(C_i = 1)$). That is, when C_i is present, the rate will be $\lambda_0 + \psi_i$, and otherwise the rate will be λ_0. This is equivalent to a continuous-time version of the Noisy-OR logic gate, used in models of discrete-time causal inference see Griffiths [181], Simma et al. [182].

We will generally assume preventative causes will thin all Poisson processes that generate effects including both the background and generative processes. One could define a preventer that only applies to some cause or a type of causal process (rather than all processes that generate the event in question) or that the preventer would have different parameter values for affecting each of them. However, I will leave such complications for further work.

A preventative cause C_j will have thinning parameter $\nexists_j$ which affects the generative processes if and only if the preventative cause is present at the time that the generative process produces its effects. Thus, if $\lambda_{\text{total}}(t)$ is the total rate, when C_j is absent, the rate be $\lambda_{\text{total}}(t)$, but when C_j is present the rate will become $\lambda_{\text{total}}(t)(1 - \nexists_j)$. This is equivalent to a continuous-time version of the Noisy-AND-NOT logic gate, which in the discrete-time setting defines the probability that an event will be canceled when the cause is present.

WHY PREVENTERS ARE NOT NEGATIVE ADDITIVE PROCESSES. You might consider treating a preventer as something that can reduce the rate of a Poisson process additively by contributing a negative rate. However, it is not clear how this would work when one considers the way it interacts with our emission model and the algebra of stochastic point processes.

Poisson processes are defined to have a weakly positive rate at all times. I use the qualifier weakly because at some times the rate could be 0, but during that time the process effectively "does not exist" in the sense that it cannot produce events at those times. It is even less clear what how to handle processes with negative rates. Poisson processes with negative rates could easily occur with negatively-valenced additive processes. All you need is for there to be a region during which the effect process had o rate and the preventer had some nonzero negative

rate that would bring the net activity of the effect to also have a negative rate.

It is difficult to comprehend what it would mean for there to be a Poisson process with a negative rate. When one counts events, one works up from zero. Negative numbers begin to exist when one has a space in which one is changing relative to some standard and when one change is able to cancel out another change and both of those changes when measured by a common measure have equal magnitude. Negative values are meaningful when we speak of spaces with arbitrary origins like physical space, time, and money. And this does not only apply to counting numbers, but any measure(including real valued measures). For example, you cannot have a negative height. You can have a height that is smaller than some standard (e.g., preventing children from going on amusement park rides) and you could have a person who was standing upside down relative to where height would normally be measured, but the magnitude of the person's height would always be positive.

A negative rate Poisson could not simply produce "anti-events" as though it were another point process, as those would actually just be events generated by another virtual point process that has some virtual existence where its events would have some preventative relationship with the original process' events. If the superposition process of the "original" and "anti-" processes is still simple(no two events can occur at the exact same time point) and they can only interact if they coöcur, then this "anti-" process would effectively not exist. It would never interact with the process it is supposed to be countering. Worse, because we postulated it as only a virtual process, we could not observe its occurrences, because they do not actually exist but only would demonstrate something like "existence" in their role in cancelling the occurrences from the original point process.

By only considering multiplicative effects on existing point processes, we can easily interpret the role of the preventer as interrupting the occurrence of any event with some probability. The multiplier model also only can have an effect if there are events available for it to stop; this is why the barrier metaphor is useful. Algebraically, a multiplicative functional form avoids the possibility of negative base-rates as even the perfect preventer would only be able to prevent all of the events (lowering the rate of the effect process to 0). This way there is no concern about what it means to prevent more than all of the events.

3.7.4 Persistent, decaying effects: convolutions and decay distributions

In most models of causation that work in discrete time or over trials in which events are treated as occurring simultaneously, a cause can only influence an effect if and only if that cause is

present on a particular trial. This is undesirable if we are to develop a framework for continuous-time causal inference, particularly if we are to allow point causes.

Not only is it useful to track how a cause's influence changes over time, instantaneous events occur for only an infinitesimal period of time. Thus, in order for such events to have emphany effect on other variables they must be able to exert influence even after they are no longer present. Thus, we will need to characterize *decay distributions*, which define how a cause's influence on its effects changes over time.◊

While I use the term decay distribution, we do not in fact require that the changes be only negative. In fact one could imagine that the rate function followed a gamma curve, where early on the "decay" is insufficient growth where the rate has not yet reached its maximum potential. Then once reaches that maximum level it decays down. One could imagine more general decay distributions as well. However, because we focus mainly on exponential decay distributions, the simple interpretation in terms of monotonic decays will suffice.

It is ambiguous, though, how one handles delays with interval causes. My models assume that interval causes have an effect during the time that they are on and no effect when they are off. But one could imagine incorporating a simple delay that shifts the time such that the effects occur shifted half of a time unit further into the future. This is formally feasible; it just requires adjusting the input value to the Dirac $\delta()$ function that we are using to convolve with the time that the cause is on. But in that convolution picture, there seems no reason why we could not have other functions than the Dirac $\delta()$ to convolve with our causes' activation times. I delay further discussion of this topic until subsection 3.7.7.

3.7.4.1 Monotonic decays: interpreting time as a filter on processes and filters

It is worth noting that this postulates the existence of causal dependencies at arbitrary temporal distances from one another and can accommodate arbitrary functions for modulating the influence. One way to avoid the potential negative effects of that is to consider monotonically decreasing functions that can only reduce the efficacy of a cause over time. Considering an exponential decay function, generative functions would find their superposed intensity functions dwindling with distance from the onset cause as if time itself acted as a filter/preventer. Con-

◊ This notion of change over time is not meant to capture that described in Rottman and Ahn [183] where the change occurs over successive presentations of the cause at various intensities to the same individual. In those cases the actual causal relationship was presumed to change. Instead, this is change associated with a single relationship and the change in its effects temporal distance from one presentation of the cause.

sidering the same, preventative functions would find their filtering probability to reverting to 0 (resulting in a gradual increase in the rate of events as time passes). One can see this as a filtering of a filter (in that it provides an instantaneous cancellation probability on any potentially active filter), without postulating the existence of any new generative processes.

3.7.5 Continuous-time Intervention

To obtain continuous time intervention, we need to modify the notion of the do operator from Pearl [28]. We will use a manicule (☞) to represent an intervention operation, in particular a two argument manicule $☞(\cdot, \cdot)$, where the first argument is the variable being intervened on, and the second is the set of times at which the intervention is occurring which can include both points and intervals.

In standard causal Bayesian networks if you have $X \to Y \to Z$, X will be independent of Z conditional on Y's value. Accordingly you can intervene on Y with $☞(Y = y)$ and it will also render X independent of Z; you are just setting Y's value to y and so the conditions are complete. In fact because of the operation now the graph looks more like $X \quad ☞ \to Y \to Z$ where X the parent of Y is no longer connected to the graph because the ☞ reached into the system and set the value, breaking X's causal influence on Y.

Unfortunately, the situation is not so simple in continuous-time — in fact it is not simple as soon as you begin to consider time at all. Even if you were to set the value of $Y(t) = y(t)$ at the current time point, if the process had occurred at any time before that, all of the variables are likely to be correlated. In the dynamic Bayesian network (DBN) case (the discrete time analogue similar to what we describe here), it takes only the number of steps equal to the directed diameter of the time-unrolled graph (the maximum number of steps it takes to reach from every individual node to every other individual node moving along one edge at a time) to ensure that all the variables are correlated.[‖] [180] Nodelman et al. [185] show that the entanglement occurs in a wide class of continuous time graphical models after any finite amount of time τ spent in the system. To see this in a Poisson process case, allow an event on X to increase the rate at which Y occurred for at least 10 seconds into the future. And Y in turn increased the rate at

[‖] Technically this requires the minimality condition that all of the edges that are represented in the graph are necessarily there and represent some kind of probabilistic dependency. Strictly speaking it is the absence of edges in these networks that make substantive claims of conditional independence; the default is for everything to be fully connected. Interestingly this minimality which itself may have close ties to manipulability and the effectiveness of intervention Zhang and Spirtes [184] at least in the standard case.

which Z occurred for at least 10 seconds into the future. Even if we intervened on Y and held it to have no events for almost up to 10 seconds 5 seconds after X affected it, it could be that Z still has a higher rate than its base-rate due to the continued influence that was broadcast by X hich in turn increased the rate at which Z occurred.

So even if we could intervene on Y at time t that would not be enough to create a conditional independence between X and Z. However, it *would* (tautologically) render Y independent from X for the time period over which the intervention lasted. And indeed, if we could screen off one variable from another for long enough for its effects to no longer be felt then we could reëstablish this notion of independence after a certain time interval. However, point interventions will be unable to accomplish that because it will only affect the intervened on variable at time t. In that case, we can define an intervention on a variable X over a period of time τ where we set the value of X to some value over that period of time τ: ☞$(X = x, \tau)$. *

That said, independence between the nodes in the rest of the graph may not be why we want to intervene. If we merely want to induce whether their is a causal relation or what the functional form of a causal relation might be, we do not need complete independence of other variables even for the period under observation. In fact, breaking causal structure may work contrary to our epistemic aims when we are trying to understand a complex causal system in its normal state of functioning.

In many cases, all we want to do is detect that because of the intervened-on events in the process whose causal status is in question some changes can be observed in some property of the effect process in question. For that, it is often sufficient to know that each of the points of intervention were *independent of each other*, not that the consequent processes have no way of interacting with each other. This is exactly the property we need to get induction "off the ground" in most of the modelling cases that we describe.

In fact, this a more realistic notion of intervention as it actually occurs in the world. In pharmacological trials, they do not aim to isolate the unique effect of a new drug when every other potential condition is under perfect control. They wish to see how the drug will interact with the causal system in its regular functioning. A drug that shut a subsystem off from the rest of the system in which it exists would more likely be a poison than a cure. Scientists only need to know that whatever happened to occur in their experiment, that the assignment of the value of the condition was independent of the assignment of all the other conditions. That is why ran-

* Technically, we could have set X to a function of values that vary over time, such as (in the case of a 6-sided die) we turn it to face 1 for one second, 2 for one second, 3 for one second, and so on. In practice, I have not used this representational capacity so I will not explore it further here.

domisation is so useful; by using a random process to determine condition identity, you can ensure that your particular condition assignments are independent and identically distributed, even when the measures that are under you observation are not.

Thus, we can still find great use from treating events as interventions at particular points and time: $☞(X = x; t)$. In particular, this allows injecting information into a system without needing to consider the causes of that injection. By fiat, the intervention itself will be independent of the other cases of intervention, even when its consequences will need to interact with the otherwise extant processes and their consequences.

Interestingly, in cases where there are no base-rates for occurrences of some nodes as part of the semantics of our graphs, we need interventions to begin to observe *any* activity on the graph. What is also convenient is that in this case, no generality is lost in assuming that the first intervention in the system defines $t = 0$. Other interventions can occur while the consequences of this first intervention are still playing out, but they will need to have their consequent activity indexed relative to some later time $☞(\bullet, s); s > 0$ where that distance is defined relative to this initial intervention.

3.7.5.1 *The effect interval of an intervention, final events and splitting processes into independent sub-processes*

If we did have a zero-base-rate process, and we did initiate the interaction, and we are dealing with an induced point process that has only a finite number of consequent events — we can state more precisely when the consequences of that intervention end.

Consider a total process E where there was only an initial intervention $☞_0$ on node X_i. Suppose that we are at the virtual time-point t' during the course of the sampling algorithm, we have not sampled any processes due to events in E following t'. There is some probability that each of those events will be induce no processes (if they occur on nodes with no children) or that the processes they do induce all produce no events. If this occurs, the set of events E is completely defined at virtual time-point t', and we can define the final event as $max(E) = e^{fin}_{☞_0}$. Thus we can define the "effect interval" of the initial intervention as $[0, e^{fin}_{☞_0}]$.

This definition can be extended for E with multiple interventions $\{☞\}^m_1$ to be the final event that descends from any particular intervention $☞_i$ on the graph $e^{fin}_{I_i}$. We can then consider the set of final events from all interventions that occur within the "effect interval" of the initial event, and call the maximum of *that* set of final events the final event within the scope of the set of interventions $e'_{☞}$. Because no events related to the initial intervention need to be

tracked, one can effectively "reset the clock" such that the next intervention after $e'_{☞}$ can be treated as starting anew. That is that new intervention can be thought of as setting $t = 0$ for a new process. This provides a convenient mechanism for splitting one process E into two parts ($E_{<e'_{☞}}$ and $E_{>e'_{☞}}$) that can be analysed independently. This is one mechanism by which one can sample "independent trials" of the effects of interventions on the same network without leaving the semantics of the formal framework to create multiple instances of the network.

You can see this as (effectively) being what occurs in the trial structure used in Lagnado and Sloman[6] or at least in terms of my model of it as described in section 3.12.

3.7.6 Finitary Poisson Processes

A 1-dimensional Poisson process can be defined in terms of a support space S and a rate function $\lambda(s)$, for which the integral $\int_{S' \subseteq S} \lambda(s)ds$, over S', a subspace of S defines the expected number of events. One can generate samples for the process by calculating the expected number of events and then sampling event times as independently identically distributed over the normalized rate function over that interval (where normalization is dividing the rate function by its mean to ensure it integrates to 1 and is a proper probability distribution). For example, a 1-dimensional homogeneous Poisson process has $S = [0, \infty)$ and $\lambda(s) = \lambda$, where for the total support S, the expected number of events is infinite ($\int_S \lambda(s)ds = \infty$) and for a proper subset $S' = [a, b)$ $\int_{S'} \lambda(s)ds = |b - a|\lambda$. As a result, one can sample from a homogeneous Poisson process over a finite interval by calculating the expected number of events and then sampling i.i.d. from the normalized rate function, which is a uniform random variable over the interval. The homogeneous rate function only affects how many events are sampled. Because the Poisson distribution does not support an infinite mean and one cannot divide by infinity, one cannot sample a homogeneous Poisson process over the entire support space.

Because you cannot sample from a homogeneous base-rate over the entire history of your model, one solution is to sample over a finite amount, and if those events play a role in your causal history, allow them to do so by incorporating their effects into the system. If a part of your causal history disallows the proposed base-rate event to have any effects (e.g., because it was a one-shot process and it already occurred), you can simply ignore the event generated by the base-rate. However, if the base-rate event would have cancelled an event that is already included in your causal history, you will not only need to cancel that event but you will need to cancel and resample any other processes that are dependent in any way upon the occurrence of the event. You do not need to do this in the case of cancelling the base-rate as (by premise) it

does not affect itself (or it would not be constant) and it cannot have had any child events You can only do so for those parts of the process that are totally unaffected by the base-rate, and if there is a directed path leading from the base-rate event to a process that can be affected by some element on that path, then that process is affected by the base-rate.

In contrast, a non-homogeneous Poisson processes (i.e., where $\lambda(s)$ varies over time but is not dependent on its own events) can have a rate function with a finite expectation over an infinite support space, $\int_{S=[0,\infty)} \lambda(s)ds = K, K \in R^+$. We call such a process a Finitary Poisson Process. We can generate samples for the entire process by first generating a Poisson random variable with mean K ($k \sim \text{Pois}(K)$), and then sample k i.i.d. random variables using the normalized rate function as the exact timing distribution. At first glance, this suggests we can define the event distribution as $P(t_1, t_2, \dots, t_k, k)$. But, because k could equal 0, there is always some probability $P(k = 0) = e^{-K}$ that no event occurs. Furthermore, the inter-arrival distribution will be dependent upon the sampled value k and will depend on which two events the interval is asked about. Nonetheless, if at least one event occurs, we can generate a closed form for the distribution of this first-event, which is the probability that *all* of the k i.i.d. events with distribution $\lambda(s)$ have *not* occurred before time t_1: $P(t_1|k \geq 1) = (1 - \int_{0,t_1} \lambda(s)ds)^k$. This allows avoiding the rejection sampling approach for nonhomogeneous Poisson processes used in Rajaram et al. [186].

3.7.6.1 Finitary Poisson processes on directed graphs

Let us consider set of N variables $\{X_1, \dots, X_N\}$, and a directed graph over those variables, where each variable is a node and edges from node X_i to X_j are encoded as (X_i, X_j) or $X_i \rightarrow X_j$. Thus the directed graph (V, E) with N nodes and and F edges can be written as

$$(\{X_1, \dots, X_N\}, \{f_{(X_i,X_j)} | f \in \{1, \dots, F\}, F < N^2; i, j \in \{1, \dots, N\}\}).$$

We will abbreviate $f_{(X_i,X_j)}$ as $f_{i,j}$ to represent the edge from X_i to X_j, including self directed edges (where $i = j$).

Each edge $f_{i,j}$ has associated with it a rate function $\lambda_{f_{i,j}}()$ that describes a finitary Poisson process generated on X_j by an event occurring on X_i (e_i). Because this is a finitary Poisson process, we know $\int_0^\infty \lambda_{f_{i,j}}(s)ds = K_{i,j}, 0 < K_{i,j} < \infty$. Thus for every e_i occurring at t_{e_i}, we can sample the entire history of each induced process $E_{i,j}$ for each child X_j of X_i, by first generating the expected number of events for $E_{i,j}$ process, and then generating the time for each event on the process at $t_{e_i} + \tau, \tau \sim \frac{\lambda_{f_{i,j}}(s)}{K_{i,j}}$. By generating every event in this way, we

know that each induced process $E_{i,j}$ is independent of other processes conditional on the time of its parent event's occurrence e_i and the rate function for that process $\lambda_{f_{i,j}}()$.

By using forward sampling in this way, we can use this independence and the superposition property of independent Poisson processes to build a single process E_{X_i} for each node X_i in the graph. Furthermore, because the processes for all of the nodes are all independent of each other conditional on the exact occurrences of each of the events, the nodes processes can be superposed to create a single process E. As a result, each event e in the total process E, can be traced back to the node X_j that it occurred on (e_{X_j}) and the particular event e_{X_i} on the parent node (X_i) induced the process which generated it ($e_{X_j} \in E_{i,j}$).

3.7.6.2 Generalized Finitary Point Processes.

One can generalize this further to Finitary Point Processes by altering the distribution to generate the number of events that occurs but generating the event times in the same way. Rather than using a Poisson distribution, one can generate the number of points from any distribution that has support over $\mathbb{Z}^+$. All of the independence properties needed are ensured by the construction of the sampling procedure. However, it is less clear how this kind of approach will generalise to the more traditional statistical account of stochastic processes (given that they will not be able to have the same kinds of analytical independence properties that standard Poisson processes have).

3.7.7 One interval cause

To deal with causes that apply over an interval of time, we can use convolution to aggregate the cause's total influence. The simplest process to convolve with an interval in time is a Dirac $\delta(\cdot)$ measure, also known as the "unit impulse". The $\delta(\cdot)$ infinite at 0 and 0 everywhere else, but integrated over the real numbers integrates to be equal to 1. Given a particular parameter describing how many events are expected to occur in 1 unit of time, this allows calculating the effective rate over an arbitrary interval of time as merely being product of the magnitude of that time interval (in time units) and the parameter in question. If this is a generative cause, we can use superposition and this will act as if the base-rate λ_0 were a greater by an additive factor ψ. For those intervals when the cause is active the expected rate would be $\lambda_0 + \psi$. If this is a preventative cause, we can use filtering and this will act as if the base-rate λ_0 were filtered by a multiplicative factor $\nexists$ for those intervals when the cause is active. For those intervals when the cause is active the expected rate would be $\nexists\lambda_0$.

If we wanted to complicate this, we could convolve the interval of time on any member of a class of functions with non-zero support on more than a point — that is, almost any function but the Dirac function (e.g., boxcar, Gaussian or Exponential curves). By convolving on functions that consider more than just the instantaneous activity rate, this would allow treating the causal influence as accumulating over the period of time when the cause is active. Let us call this an aggregative interval cause.

This contrasts with the alternative stative approach, where all that matters is whether the cause is on or off (as is described in subsection 3.7.4). One could still take into account delay functions in the case of stative interval causes, but doing so is somewhat more challenging (or at least not uniquely determined). For example, you could consider an induced Poisson process with a rate that begins at the onset of the cause and whose effects held at a constant for the duration of the cause, but whose effects decay according to a decay distribution after the cause was turned off. This is equivalent to convolving by the Dirac δ for the time when the function is on and then triggering a point cause at the moment that it turns off.

One could also treat the cause onset as a point cause in its own right. This would allow influence to stretch forward in time after the instantiation of onset as though it were a point cause, but to still modify it based on the properties of the interval. This would not necessarily rely on convolution✦. But, for example, you could imagine that the effect decays but lasts only until the moment when the cause turns off (as though one opened a tap to a finite keg filled with marbles that flow out at a decaying rate until the tap is closed). This could be accomplished in the same way as a preventative version of the above stative case (where the OFF state switch induces a point event cause) or the perfect preventative interpretation of the process governing one-shot events. In either case, one would treat the event of switching off as a perfect preventer of the process in question for all times going forward while that cause is off.

Yet more complicated constructions can be found, for some ideas of what those can range over, see subsection 3.14.2.

Regardless, if one does not know the time at which the events occurred, all one needs to

✦ One actually *could* see point causal influences in terms of convolutions. But it requires changing the way in which convolution is used. Instead, one would convolve a decay function with with a identity function that is true at all times after the initial event occurs. Then this produces the appropriate decay function as each subinterval will be given the intensity measure that accords with the integral initiated at that time. This has the advantage of only postulating simultaneous causal influences, but at the cost of proposing the existence of an infinitely long state of affairs for every event that occurs with this kind of decay distribution. It also suggests ways of composing the relationship with other relationships, for example, preventers that can be seen to simultaneously occur and interfere with the continually existing generative causal process.

know is the magnitude of the intensity measure for the period of time under observation, meaning that intensities influenced by decay functions are indistinguishable from constant, stative or aggregative function that for any one period of observation, happen to have the same intensity measures.

3.7.8 One-shot events

A one-shot event is a process for which the first arrival is also guaranteed to be the last arrival for that process. This would apply to events that in the history of an entity can only occur once such as death. This is in contrast to events that can occur multiple times, such as the act of breathing. For events of this type we have a fairly standard likelihood form. For a one-shot event occurring at time t it is the product of the probability of observing nothing according to the model's rate function ($F(N(0,t) = 0|\lambda(\cdot))$), the instantaneous rate of occurrence at the time of occurrence ($\lambda(t)$), and the probability that nothing occurs after that for the rest of the observation period (which by definition of a one-shot cause, is 1 no matter how long the observation period).

In loglikelihood form this is:

$$\begin{aligned}\ell(N([0,t),(t,\infty)) = 0, N(t) = 1|\lambda(\cdot)) &= \log(F(N(0,t) = 0)) + \log(\lambda(t)) + \log(1) \\ &= \log(\frac{\Lambda([0,t))^0}{0!}) + \log(e^{-\Lambda([0,t))}) + \log(\lambda(t)) \\ &= -\Lambda([0,t)) + \log(\lambda(t)). \end{aligned} \tag{3.7}$$

The loglikelihood that no event happened during the observation period($[0,u)$) is:

$$\begin{aligned}\ell(N([0,u) = 0, |\lambda(\cdot)) &= \log(F(N(0,u) = 0)) \\ &= -\Lambda([0,u)). \end{aligned} \tag{3.8}$$

In the case where one knows an event happened at some time in an interval (as in section 3.10) we need to integrate over this probability for all the values of t in that interval. In loglikelihood

form this is:

$$\begin{aligned}\int_{t_1}^{t_2} \ell([0,s))ds &= \int_{t_1}^{t_2} \log(F(N(0,s)=0))ds + \int_{t_1}^{t_2} \log(\lambda(t)) \\ &= \int_{t_1}^{t_2} -\Lambda([0,s))ds + \int_{t_1}^{t_2} \log(\lambda(t)).\end{aligned} \tag{3.9}$$

Foundation for Causal Ascription: Law. Some processes are one-shot processes, meaning that to capture causal systems with this feature, we will need to accommodate one-shot causes. But that is not the only reason to study one-shot events. One-shot events are important as they are the basis of many cases in which we use causal ascriptions (especially legal settings). Though the law acknowledges that there can be multiple causal contributions to the same event, it nonetheless is organised around assigning fault and liability to "counts" of individual events. However, one cannot assign blame on the basis of a causal system one knows nothing about, meaning that the ascriptions need to rest on background beliefs inferred from previous instances of events like those under analysis. Causal induction is exactly the kind of belief capable of supporting those claims, but induction(and any other statistical inference procedure) faces specific difficulties in cases with one-shot events.

3.7.9 Statistical inference in one-shot processes.

Any one instance of a one-shot process (a process defined by eventually producing a one-shot event) can at most provide a single sample for a causal learning task. In the case that the causes of that process are stochastic in nature, one cannot infer statistical knowledge (including knowledge obtained via causal induction) without reference to more than one instance of the event. We need ways of obtaining multiple samples in order to collect enough data to be able to induce causes in stochastic systems.

One way to obtain multiple samples is to have many instances of a process, and have that process only truly occur once. This describes the data available in the studies by Greville and Buehner [2]: 40 bacterial cultures that do or do not die on particular days, they can only die once. Each bacterium in one experimental condition is assumed to be identical and independent conditional on their causal influences. This allows multiple samples of the process, providing a statistical/distributional dataset of whether death occurs and the amount of time until the deaths that do occur. This can be thought of as a way to use the ontology plus the facts of the world to provide multiple instantiations.

Another way of obtaining multiple samples allows multiple trials on the same system where the "one-shot" property only holds for the extent of a trial. One can see this as the system returning to an equilibrium state from which it may be perturbed, though that perturbation is uncorrectable for the course of a trial. Different trials are rendered independent of each other by virtue of the equilibrium state. The persistence of entities in the system across trials allows the identification of cross-trial activity to be seen as multiple instantiations, providing the sort of statistical evidence needed for causal induction. In fact, this is one way of caching out the *meaning* of a trial in continuous time. Trials of this sort – where events internal to one trial can occur at different times, but where each event can occur only once in each trial – can be seen in Lagnado and Sloman [6]. Lagnado and Sloman [6] have four computers that exhibit behaviour that is the consequence of the same unknown network structure across 100 trials either do or do not become infected with a virus sent from one of the computers.

3.7.9.1 Alternative view: perfect prevention

There is an alternative view to one-shot events that, though formally equivalent to the "first arrival is last arrival" property, has different implied compositional semantics. In this picture, a one-shot event is a process that, upon the arrival of the first event induces a perfect preventative process on itself. A perfect preventative process is just a preventer that is always present, has no decay in its effect and its $\nexists = 1$.

The key difference is that because the preventative one-shot property is defined using the introduction of a new causal process, that causal process could interact with other causes in the system. Those other causes could eliminate the original process' "one-shot" property by interfering with the preventative mechanism that is producing the effect. This is analogous to the immunity one acquires after surviving some diseases. After you first are infected with the chicken pox, assuming you recover, you should not be able to become sick from the chicken pox again. But if your immune system were to be suppressed (e.g., by taking methamphetamine [187]) it is possible that this preventative system would itself be prevented from enacting its usual effects.

3.7.9.2 Triggered events

If we consider Mackie's [149] "coin-slot" machines (see minor section 3.3.2.1), the most basic version produces one outcome (a chocolate bar) for each coin that is inserted. The system responds with exactly one event to each instance of a triggering cause. From a process perspec-

tive, we can think of each trigger event as producing a new process, that happen to be assigned to the same object. This is closely related to the framework used by those who coerce continuous time phenomena into contingency tables. This also seems to be the causal notion that plays into people's ideas of particular events having particular, unique causal events (e.g., when assigning fault due to a car crash).

The idea of pairing particular causes to particular events becomes challenging when events like the event in question can occur multiply within the same process. For example, a triggering cause that produces two or more events cannot be easily understood through a contingency table, especially if those events delayed and it is possible to trigger the mechanism again before the delays have completed. The more the wait-times between causes and their effects overlap the more difficult the problem of particular assignment will become.

3.7.9.3 Multiple causes of a one-shot event.

It is possible that there could be many contributions to a one-shot process, as is the case in death caused by chronic environmental exposures. Even though there is a one-shot process One could see this as one reason that occupational diseases are more difficult to assign particular fault to than workplace injuries. Though both occur at work, the occupational diseases tend to have multiple instances of the cause making it difficult to assign particular blame to any one event, whereas workplace injuries result from individual events with unique effects that could not easily be attributed to other causes.

The extreme version of this contrast is that of a interval cause rather than a point cause. If one imagines more and more instances of a multiply contributing cause while maintaining the same influence, one arrives at the notion of an interval cause. Given the difficulty of assessing the effects of an aggregative versus a static interval cause on multiple events occurring at unknown times(see subsection 3.7.7), it is unsurprising that one effect happening at a known time would be difficult to link to an interval cause. This is made worse if there is a long delay between the occurrence of the cause. If it is an aggregative interval cause, was it the delay relative to the duration that mattered? Or if it was a static interval cause, was it the onset that mattered?

Considering multiple causal contributions to a one-shot process is necessary in cases where you believe there is an unknown mechanism that allows for such multiple contributions. In the Lagnado and Sloman [6] experiment, we do not know which of the computers are potential causes of which others, and any one could in actuality have multiple inputs. In that case, we

need to be able to define how the influences interact, and in section 3.12 we consider only additive, generative causes.

3.7.10 Defining a rate with multiple independent causes

In the case where there are many causes, if we assume they are (at least conditionally) independent, and they have the generative and preventative forms under discussion, their effects can be composed with one another. In that case, such that you will have a summation of the $\psi_{\cdot}$ rates for the generative causes and a product of $1 - \nexists_{\cdot}$ the thinning parameters for the preventative causes. The rate function for a case with background rate λ_0 and i generative causes and j preventative causes is defined as

$$\begin{aligned}\lambda(t) = &\Big(\lambda_0 + \sum_i \psi_i \int_{T' \in T_i(C_i=1)} \delta(t, T')dT'\Big) \\ &\prod_j (1 - \nexists_j \int_{T' \in T_j(C_j=1)} \delta(t, T')dT').\end{aligned} \tag{3.10}$$

where $\delta(\cdot, \cdot)$ is the 2 argument Dirac delta function (see minor section 3.5.8.4).

This does not include a prior for the causes; we can mathematically justify this if we treat these cause instances as continuous-time interventions.

3.7.11 Decomposing likelihoods on the basis of point events

One of the key insights needed to have a framework for that addresses causal relations between point processes is recognising that you can splice the sequence of data into arbitrary periods of time. Notably, this includes retroactively splicing observation sequences based on when those events occurred. We do this primarily for computational and analytical convenience but there is a conceptual basis as well.

The most central assumption is that causes can only propagate effects into the future, i.e., that there is a fundamental asymmetry in our notions of cause and effect. We can think of the total history of the world (in my case, a set of stochastic processes) as capable of contributing to its current state of affairs, that means that every time an event occurs history changes and thus the manner of determining the current state of affairs is slightly different. That means that we can define each interval of time between events in terms of the available causal influences at those times. We then can splice time into those intervals during which no events occurred and separately compute the likelihood of each interval taking into account what event occurred

at the end of the interval. If a state changes, that change point can act as an event that factors into how we divide time into intervals. If a variable with a real value changes smoothly, we can address those changes using integrals(for point and state effects) and differential equations (as they affect other real values). If a real-valued variable has a discontinuous change, we can treat that discontinuity itself as a point event and will splice the interval containing the discontinuity into two intervals: one ending at and one beginning at the discontinuity[⇓].

If we presume that we have observed all of the events relevant to our system, then we can describe the total probability of a set of point events as the product of the historically conditioned sequence of time intervals. That is, because we know the total history of all the potential influences on the system at the occurrence of every point and over every interval between the points. We know all of the influences over every interval between points because (by definition) we know that nothing else occurred during that interval, meaning no new information could be introduced into the causal history determining every other event.

3.7.11.1 Splitting a set of point events

In order to deal with the results of multiple cause point interventions, it will be useful to look at a series of point events as a sequence of intervals during which nothing happened and those points at which events did occur. This requires having fully observed all of the events in question in order to have well formed conditional likelihoods and conditional intensity functions.

What may be counterintuitive is that in order to make this computationally well formed, we will be splitting the total event set, that is we will be splitting not only using the effect events, but based on the cause events. If there were only one causal point event (ς) at the beginning of the event series (as in section 3.10), we can analyse this only by splitting the effect event set into a series of intervals between the effect events and the instantaneous points at which the effect occurred. This would then require knowing the measures for those intervals and the instantaneous rates at the time at which the events occur. That is, if we intervened at time 0 making $\varsigma_1 = t_0 = 0$ (where no effect occurred), the effect set is $\{t_1, \dots, t_k\}$ with observation stopping at t_{fin}.

Note, we need to remove the $\ell(N(t_0) = 1)$ because it is counted in the sum in order to give it nice indexing properties, when in actuality it should not be, as no effect event occurred there and we assume our causes are point interventions.

[⇓] This approach will not work for poorly behaved function such as the Cantor set or the Dirichlet function (which takes on values of 0 at all irrational numbers and 1 at all rational numbers). Fortunately, we deal with analytical functions

To see why we need to split time on both effect and cause events consider the case where one cause ς event occurs between two effect events t_1, t_2. For the sake of concreteness, let us say that the cause is a generative cause. During the time period $[t_1, \varsigma)$, the non-occurrence of events only needs to be explained with reference to the intensity function as determined by all influences (including the base-rate) that occur before the cause event at ς. However, after the cause occurs, the event generating process operates under a different intensity than before. In particular, now the intensity measure needs to consider the influence of the cause that occurred at ς. If the influence of the base-rate is constant and the influence of a cause is monotonically decreasing, and the cause is generative, $1s$ of no events occurring after the cause is less likely than $1s$ of no events occurring before the cause event, or supposing that $|\varsigma - t_1|, |t_2 - \varsigma| > 1$ then $\ell(N(\varsigma - 1, \varsigma) = 0|\Theta) > \ell(N(\varsigma, \varsigma + 1) = 0|\Theta)$.

Most importantly, it will be easiest to define those loglikelihoods separately rather than defining the loglikelihood over the total interval $\ell(N(\varsigma - 1, \varsigma + 1) = 0|\Theta)$. This occurs because there is a discontinuity in the rate exactly when the cause occurs ς, meaning that the intensity cannot be integrated without addressing this discontinuity.

This complication is why we define our intervals in terms of sets that are closed on the left and open on the right. This(plus the decay measures we consider) results in our intensity functions being càd-làg functions. That means that by splitting our loglikelihood at the points of discontinuity, we will be able to integrate our intensity function using a smooth underlying function for each integral.

The observation period (which spans $[t_0, t_{\text{fin}})$)[•] can extend beyond any of the observed events. So if we modify the above to consider splits that occur on ςs, the observation period, and the to-be-explained events, we will need to consider the sequence of events composed of the sorted union of the observation bounds, the cause times, and the effect times ($\mathbb{t} \stackrel{\text{def}}{=} \text{sort}(\{t_0, t_{\text{fin}}\} \cup \{\varsigma_1, \dots, \varsigma_n\} \cup \{t_1, \dots, t_k\})$). For notational convenience, will need to consider the set of intervals in question $\{\tau\} = \{(\tau_1, \tau_2)_j\} = \{(\tau_{j(1)}, \tau_{j(2)})\} \stackrel{\text{def}}{=} \bigcup_{t_i, t_{i+1} \in \{\mathbb{t}\}} (t_i, t_{i+1})$, where the number of intervals will be equal to $n + k + 1$ (the number of causes, the number of effects and the observation boundaries minus 1 because they are intervals). This loglikelihood

[•] Note, t_{fin} is not the last observed event (t_k), but the final time of observation.

then will look like

$$\ell(\{\mathbb{t}\}|\{\varsigma\}_1^n, \{t_0, t_{\text{fin}}\}, \Theta) = \sum_{j=1}^{n+k+1} \ell(N(\tau_{j(1)}, \tau_{j(2)}) = 0 | t \in \mathbb{t} \forall t < \tau_{j(1)}, \Theta) + \lim_{\Delta \to 0} \ell(N([\tau_{j(1)}, \tau_{j(1)})) = 1 | t_\cdot < \tau_{j(1)}, \Theta), \tag{3.11}$$

where the particular form of $\ell(\cdot)$ will depend on the identity of the elements in each interval $(\tau_{j(1)}, \tau_{j(2)})$.

If $\tau_{j(1)}$ is a cause (ς) or an observation boundary, these are taken as given (causes and observation periods are treated as effective interventions) and so the $\ell(N(\tau_{j(1)}) = 1) = 0$ (remembering that a 0 as a loglikelihood is a 1 as a likelihood, making these events certain).

If $\tau_{j(1)}$ is a effect event then $\ell(N(\tau_{j(1)}) = 1|\cdots) = \log(\lambda(\tau_{j(1)}|\cdots))e^{\Lambda(\tau_{j(1)}, \tau_{j(1)})} = \log(\lambda(\tau_{j(1)}\cdot$

By construction, all intervals contain no events, and so all of the interval loglikelihoods can be treated identically ($\ell(N(\tau_{j(1)}, \tau_{j(2)}) = 0) = \log(\frac{(\Lambda(t_{j(1)}, t_{j(2)}))^0}{0!}) \exp(-\Lambda(t_{j(1)}, t_{j(2)})) = \log(1) - \Lambda(t_{j(1)}, t_{j(2)}) = \Lambda(t_{j(1)}, t_{j(2)})$).

This is enough to provides the log-likelihood once we have conditionally defined $\lambda()$ and $\Lambda()$ functions.

3.7.12 One rate to rule them all: A total rate function with — Multiple causal instances; Generators and Preventers; Instantaneous and Interval causes; Decay functions

Let us suppose we want to take into account the possibility of everything that I have been discussing: for example, that there are many potential generative and preventative causes, that these causes can exist as both instants and intervals, that intervals can take on many forms with and without decay functions (aggregative, stative, &c.), and that these points and instances can occur multiple times. To do so is somewhat complicated, but feasible if we step through the complexity slowly.

Let $\dot{f}_i^g(\cdot, \cdot; \dot{\gamma}_i)$ indicate a decay function with unknown parameters $\dot{\gamma}_i$ for point generative cause C_i, and $\dot{f}_j^p(\cdot, \cdot; \dot{\theta}_j)$ indicate the decay function with unknown parameters θ_j for preventative cause C_j. We can let this same set ($\dot{f}_i^g(\cdot, \cdot; \dot{\gamma}_i)$ and $\cdot f_j^p(\cdot, \cdot; dot\theta_j)$) indicate any cause that effectively acts as a point cause even if it is derved from an interval. Though formally

somewhat distinct, these can just be included in the same set of point events. This would include intervals that only introduce an effect when they initially occur. This is true even if they are canceled when the event no longer occurs although in that case, they are actually a conjunction of two causal events, one generative and one preventative, where the preventative cause is perfectly preventative and does not decay. In fact, we can *also* use them to handle the "stative" causes that experience no decay while active but have decaying effects after they are turned off — in that case, we treat the interval portion as described in the next section on interval causes and the point at which the state "turns off" as a point cause that is added to the set of point causes with its own decay parameters.

For stative interval causes, let $\bar{f}_i^g(\cdot, [\cdot, \cdot); \bar{\gamma}_i)$ indicate the convolution of the decay function with the Dirac $\delta()$ function with unknown parameters $\bar{\gamma}_i$ for generative cause C_i, and $\bar{f}_j^p(\cdot, [\cdot, \cdot); \bar{\theta}_j)$ indicate the aggregation function with unknown parameters θ_j for preventative cause C_j. We use the bar ($\bar{f}$) notation in order to convey that this is a stative interval, not a point cause for which we use the dot ($\dot{f}$) notation, nor is it a aggregative cause, to which we give a hat ($\hat{f}$) to indicate its gradual increase over time that it is active.

For aggregative interval causes, let $\hat{f}_i^g(\cdot, [\cdot, \cdot); \hat{\gamma}_i)$ indicate a aggregation function with unknown parameters $\hat{\gamma}_i$ for generative cause C_i, and $\hat{f}_j^p(\cdot, [\cdot, \cdot); \hat{\theta}_j)$ indicate the aggregation function with unknown parameters θ_j for preventative cause C_j.

Let the set $\{\varsigma\}_1^{n_\iota}$ be the set of n_ι times that instantaneous cause C_ι occurs, $\{\bar{\tau}\}_1^{m_\iota} = \{[\bar{t}_\uparrow, \bar{t}_\downarrow)\}_1^{m_\iota}$ be the set of m_ι stative intervals (each of which starts at $\bar{t}_\uparrow$ and ends at $\bar{t}_\downarrow$) over which interval cause $C_\iota^{\bar{\tau}}$ occurs, and $\{\hat{\tau}\}_1^{a_\iota} = \{[\hat{t}_\uparrow, \hat{t}_\downarrow)\}_1^{\nu_\iota}$ be the set of a_ι aggregative intervals (each of which starts at $\hat{t}_\uparrow$ and ends at $\hat{t}_\downarrow$) over which interval cause $C_\iota^{\hat{\tau}}$ occurs, where ι is replaced with the index (i or j) of a particular generative or a preventative cause. We set $\lambda_\varnothing$ to indicate the underlying rate of an effects' occurrence. Let $\dot{\psi}_i$ and $\dot{\nexists}_j$ to indicate the maximum value of a point cause's influence, $\bar{\psi}_i$ and $\bar{\nexists}_j$ be the maximum value of a stative cause's influence, and $\hat{\psi}_i$ and $\hat{\nexists}_j$ be the maximum value of the aggregative causes influence. Finally, we will consider the set of generative causes $\mathcal{C}_g$ and the set of preventative causes $\mathcal{C}_p$, though these identify the causal processes not the entities which allows the same entity to potentially

produce a generative and a preventative causal process.

$$\lambda(t) = \Bigg(\lambda_{\varnothing} + \sum_{C_i \in \mathcal{C}_g} \Big(\sum_{\dot{\varsigma} \in \{\dot{\varsigma}\}_1^{n_i}} \psi_i \dot{f}_i^{g}(t, \dot{\varsigma}; \dot{\gamma}_i) + \sum_{\bar{\tau} \in \{\bar{\tau}\}_1^{m_i}} \psi_i \bar{f}_i^{g}(t, \bar{\tau}; \bar{\gamma}_i) + \sum_{\hat{\tau} \in \{\hat{\tau}\}_1^{\nu_i}} \psi_i \hat{f}_i^{g}(t, \hat{\tau}; \hat{\gamma}_i) \Big) \Bigg) \times \prod_{C_j \in \mathcal{C}_p} \Big(\prod_{\varsigma \in \{\varsigma\}_1^{n_j}} (1 - \dot{\eth}_j f_j^{p}(t, \varsigma; \theta_j)) \times \prod_{\bar{\tau} \in \{\bar{\tau}\}_1^{m_j}} (1 - \bar{\eth}_j \bar{f}_j^{p}(t, \bar{\tau}; \bar{\theta}_j)) \prod_{\hat{\tau} \in \{\hat{\tau}\}_1^{\nu_j}} (1 - \hat{\eth}_j \hat{f}_j^{p}(t, \hat{\tau}; \hat{\theta}_j)) \Big). \tag{3.12}$$

3.8 Causal induction using continuous-time causal theories

3.8.1 Inferring functional form: rates, tables & in continuous time

The first kind of problem I will analyse with the CTCT framework is that of inferring the form of an elemental causal relation, given that a relationship is present. I consider the functional forms of event generation and event prevention by using the superposition and thinning operations described in the formal framework. The problem of inferring the form of an elemental causal relationship is defined in terms of using an approximation to Bayesian inference with a prior over potential graphs and the implicit parameterisation for those graphs. I draw samples from that prior in order to compute the necessary arguments to define the likelihood that connects the parametrised graph to the observed data. After this I normalise the graph-specific likelihoods, and – combined with the edge-set prior distribution – compute a posterior distribution over the graphs. With the posterior distribution over the graphs (which have different functional forms), I can compute a prediction for a normative judgements. I optimise the scale of those judgements over the entire set of conditions included in the experiment to produce the best correlation with average human judgements on a variety of measures.

Functional form inference of inference covers section 3.9 (where I model human judgements formed on the basis of rates sampled from uniform periods of time), section 3.10 (where I model human judgements on the basis of tabular data where events occur at most once at some unknown point within one of five days), and section 3.11 (where I model human judgements on the basis of continuously observed sequences of events, with multiple cause instances and multiple effect instances). Human judgements in the first case will be based on a 0–100

scale stating the degree to which they were confident in the existence of a cause of a particular type(generative or preventative). In the second case, human judgements were given along a single scale ranging from -100 to 100 where more negative scores indicated a stronger belief that a cause was preventer and more positive scores indicated a stronger belief that the cause was a generator. In the last case, human responses were given on a proper simplex — a categorical probability distribution with generator, preventer and the null graph (where the cause has no effect) as the three categories. In all cases, *average* human judgements on these measures are the aim of my model fits.

3.8.2 Inferring causal structure: trials, hidden mechanisms & streaming data

The next kind of problem I analyse has two flavours but fundamentally comes down to inferring the causal structure that generated a sequence of events. This problem can still be defined by means of analysing a posterior over graphical models, however I will have many more than two or three models to work with. I similarly will define a prior for different parameterisations of graphs and will marginalise over samples drawn from the prior distribution to provide the means of explicitly defining the likelihood. These likelihoods happen to be more complicated than those in the functional form problem (possibly excepting the model in section 3.11) because they involve many cause occurrences and many effect occurrences. In fact, in section 3.12, I deal with the case where hidden causes may be both causes and effects of each other. Nonetheless, the final inference is formed by considering the variety of causal graphs, computing the posterior over those graphs and computing statistics about those posteriors. In most cases, those statistics are summary statistics about the presence of edges shared across graphs because they share the same parent and child nodes.

3.9 Inferring form from rates

3.9.1 Semmelweis and puerperal fever

> *Everything was in question; everything seemed inexplicable; everything was doubtful. Only the large number of deaths was an unquestionable reality.… None of us knew we were causing innumerable deaths.*
>
> Ignaz Semmelweiz, 1861 [188]

Between 1841 and 1847 Vienna General Hospital's First Obstetrical Clinic had two wards in which women would undergo child-birth: the midwives' ward and the doctors' ward. Aside from the staff, the wards were extremely similar barring one important feature: women admitted to the doctor's ward would have twice the chance of dying from puerperal fever (also known as childbed fever) than those admitted to the midwives ward. In fact, women who were unable to reach the hospital in time and delivered in the streets would have a better chance of surviving child-birth than if they were to be admitted to the doctor's ward. This situation was distressing, and the hospital's solution was to rapidly remove women suffering from puerperal fever from the doctor's maternity ward to the general hospital so the difference between the figures would seem less stark (these relocations were not performed in the midwives ward). This solution was unsatisfactory to at least one person, and, accordingly, Ignaz Semmelweis made it his duty to try to isolate the cause of the malady.

Semmelweis noted that the methods of the two clinics were the same, and that other diseases did not differ in the rate of affecting the two wards. He ruled these causes out. He noted additionally that newborns and mothers would die from the same disease, suggesting that it could not be a sickness specific to trauma from the act of labour (as the infants had not given birth). This was important as the commonly held theory (derided by Semmelweis) was that the disease itself (and the death that resulted) was due to the combination of the enlarged state of the postpartum uterus (the puerperal state) and the fear of death (presumably by puerperal fever). In addition, the faculty of medicine decided that male foreigners studying in the doctoral ward were to blame as they were too rough with the patients. They were expelled from the hospital, and as Semmelweis noted the rates of death in the two clinics were unaffected. He found no difference in the location within the ward all were equally affected. He changed his delivery practice to have the women lie on their sides rather than on their backs, it made no difference.[188] They altered the priest's path through the ward when giving last rites to they dying, so as to be more secret and not remind women of the death they fear and by so doing instil and cause it. Needless to say given todays understanding, rerouting a minister's walk did not affect the death rate in the ward.[189]

In response to this, the problem was declared to be the result of "atmospheric-cosmic-terrestrial" conditions of Vienna, an "epidemic" and therefore impossible to affect by intervention. The administration "did not consider the purported cause but only the number of cases… many patients became ill and died, [so] it was identified as an epidemic." And as a result "the unfortunate confusion between the concepts of epidemic and endemic disease delayed discovery of the true cause of childbed fever."[188]

But Semmelweis did not give up. He became convinced that it was because doctors were transferring cadaverous matter from the autopsy rooms to women in the maternity ward and thereby causing the fever. To address this, Semmelweis instituted a regimen of washing his hands with liquid chlorine and brushing his nails with a nail brush before beginning any obstetric activities and required that his students did the same. He instituted the practice in May 1847, when 12.2% of the women in the first clinic died of puerperal fever. By June 1847 that rate was down to 2.38%, the latter half of the year's total death rate from puerperal fever was 3.04%. In 1848, after Semmelweis began taking students aside and lecturing them on their handwashing practices, the rate for the year in the doctor's clinic was 1.27%. For a summary of the data before and after, see Table 3.1. After he left Vienna, going to Pest, Hungary where between 1851 and 1855 eight patients died of childbed fever (of the 933 births he attended, a rate of 0.86%). Semmelweis [8] had successfully identified a causal method that dramatically reduced the rate of death by childbed fever. Unfortunately, because his work was not well received by his colleagues, this causal factor was limited to follow him whenever and wherever he was present.

3.9.2 Rates as minimally informative temporal data

In the sense of carrying the least amount of information, the average rate over a time period, irrespective of when particular events occurred, is the simplest form of continuous-time data. The first analysis of the inferring functional form in this manner from rate information can be found in Griffiths and Tenenbaum [30]. Participants in their experiment were asked to determining whether the application of different electrical fields changes the rate at which radioactive compounds emit particles. The cause event is framed not in terms of a point event at which an electric field is applied, but as a continuously active electrical field, with the particles being counted in the two cases.

Such a system that can be analysed using this model, where it has one effect (particle emissions) with a background rate and (possibly) one generative cause (the electrical field, C_i). Griffiths and Tenenbaum [30] presented participants with information summarizing the number of effect occurrences (particle emissions) that occurred during one minute with the cause on ($r(E|C_i([0,t]) = 1) = \frac{\#_{C_i}(E,[0,t])}{|[0,t]|} = \frac{\#_{C_i}(E,[0,t]}{t}$), so when $t = 1$) and one minute with the cause off. For each compound, participants rated on a scale of 0 (the electric field definitely does not cause the compound to decay) to 100 (the electric field definitely does cause the compound to decay) their belief regarding whether C_i was indeed a cause.

	First Clinic				Second Clinic		
	Births	Deaths	Rate	Handwashing?	Births	Deaths	Rate
1841	3,036	237	7.8		2,442	86	3.5
1842	3,287	518	15.8		2,659	202	7.6
1843	3,050	274	9.0		2,739	164	6.0
1844	3,157	260	8.2		2,956	68	2.3
1845	3,492	241	6.9		3,241	66	2.0
1846	4,010	459	11.4		3,754	105	2.8
1847(total)	3,490	176	5.0		3,306	32	1.0
Jan–May	1,534	120	7.8		N/A	N/A	N/A
June–Dec	1,841	56	3.0	✓	N/A	N/A	N/A
1848	3,556	45	1.3	✓	3,319	43	1.3
1849	3,858	103	2.7	✓	3,371	87	2.6

Table 3.1: Absolute counts and rates for cases of pregnancy and death by childbed fever in the two maternity clinics in the Viennese General Hospital from 1841–1849. Physicians attended to patients in the first clinic and midwives attended to them in the second clinic. Until mid-May 1847 physicians would not wash their hands, brush their nails or otherwise engage in cleaning procedures between performing autopsies and obstetric examinations. There were dramatically different rates of death by childbed fever in the two wards, until Semmelweis[8] forced reforms that brought better hygiene and (accordingly) improved survival rates for women going through childbirth in the first ward. Data of this sort (derived from Semmelweis[8], rates calculated manually to avoid mistakes in the original report) are similar to those needed to induce causal relations from rates. Though we can consider the institution of handwashing in analogy to "turning on an electric field", unlike my experiments, there is no baseline number of potential cases from which events arise. No month level data is available for the second clinic.

To model the participants' predictions, Griffiths and Tenenbaum[30] treated the problem as one of model selection between a graphical model G_0 where the cause had no effect (i.e., $\psi_i(t) = 0, \forall t$) and a graphical model G_1 where the cause did have an effect (i.e., $\psi_i(t) > 0, \exists t$). They parametrised G_0 and G_1 as we have above, as Poisson processes with different rate functions, where generative causes are treated as we have treated them above. The quantity used to predict human judgements, termed "Causal Support", was the log likelihood ratio in favour of G_1, integrating over the values of all of the parameters of the Poisson process. This model performed well at predicting the mean judgements of the participants, with a scaled correlation of $r = .978, \alpha = .35$.[❄] Other models considered by Griffiths and Tenenbaum[30]

[❄] As is usual in these studies, the authors scaled their model's values with the non-linear transforma-

also performed well, with the raw difference in rates ΔR from Anderson and Sheu [190] giving $r = .899$, $\alpha = .05$, a variant on the Power-PC theory from Cheng [81] giving $r = .845$ $\alpha = .06$, and a modified χ^2 score giving $r = .980$, $\alpha = .01$.

Griffiths and Tenenbaum [30] considered only generative causes, creating the opportunity to use the same paradigm to evaluate whether the treatment of preventative causes outlined above is effective. We ran a new experiment to address this question. Considering only one preventative cause, we used nearly identical materials to Griffiths and Tenenbaum [30], only changing the word "increases" to "decreases" and using the following $(N(c^-), N(c^+))$ pairs (where $(N(c^-)$ and $N(c^+))$ are the number of particles that were emitted during the minute when, respectively, the cause was absent and was present): $(52, 2)$, $(60, 10)$, $(100, 50)$, $(12, 2)$, $(20, 10)$, $(60, 50)$, $(4, 2)$, $(12, 10)$, $(52, 50)$.

We recruited 18 participants through Amazon Mechanical Turk to participate in my study online. We asked each participant to make the following judgement about each of the nine cases: "Does this field decrease the rate at which this compound emits particles?" Participants responded on a scale ranging from 0 (the electrical field definitely does not decrease the rate of particle emissions) to 100 (the electrical field definitely does decrease the rate of particle emissions).

Following Griffiths and Tenenbaum [30] we modelled this task as a model selection problem between two graphs G_0 where the cause has no effect and G_1 where the cause has a (preventative) effect on the rate of particle emissions. We use a simplified version of the model defined in Equation Equation 3.10 with one potential preventative cause with parameter $\nexists_1$ and a background rate λ_0 to define the likelihood functions for G_0 and G_1. We assumed that in G_0, $\nexists_1$ is constrained to be equal to 0. To obtain the log likelihood ratio, we need to provide likelihoods in terms of the graphical models (i.e., $P(D|G_0)$ and $P(D|G_1)$ for the observed data D). However, as they stand, the Poisson processes associated with these graphical models assume that the parameters λ_0 and $\nexists_1$ are known, which is not the case. We thus need to define prior distributions over these parameters. With defined prior distributions, we can use Monte Carlo integration to obtain marginal likelihoods, corresponding to the probability of the data given just the graphical model. We defined the prior for $\nexists_1$ as $U(0, 1)$, i.e., uniformly distributed in the interval $[0, 1]$. Griffiths and Tenenbaum [30] used an improper prior for λ_0, with $\lambda_0 \sim \frac{1}{\lambda_0}$. We approximated the previously used prior by sampling $v_0 \sim U(\log(10^{-6}), \log(10^6))$ and letting $\lambda_0 = e^{v_0}$. [*]

tion $y = \text{sign}(x)|x|^\alpha$ where α is chosen to maximize the linear correlation r.

[*] To see the approximation, note that $v_0 = \log(\lambda)$ and $v_0' = \frac{1}{\lambda}$ and use a change of variables to find

Using the log likelihood ratio in favour of the hypothesis that $\nexists_1 \sim U(0,1)$ (G_1) over $\nexists_1 = 0$ (G_0) as the predictor of mean human judgements, there is a high scaled correlation with the results of the experiment: Causal Support gives $r = 0.963, \alpha = 0.23$ (see Figure Figure 3.7). I also evaluated the models tested by Griffiths and Tenenbaum [30], which showed similarly high performance, $\Delta R : r = 0.780, \alpha = 1.95 \times 10^{-4}$; Power PC: $r = 0.986, \alpha = 0.45$; and $\chi^2 : r = 0.942, \alpha = 1.95 \times 10^{-4}$. My purpose is not to claim that the model I have defined is the best model of human inference, but to demonstrate that the assumptions I have made about handling preventative causes in my framework are reasonable. Future work will hopefully clarify whether this model outperforms the other models in cases where their predictions diverge more dramatically.

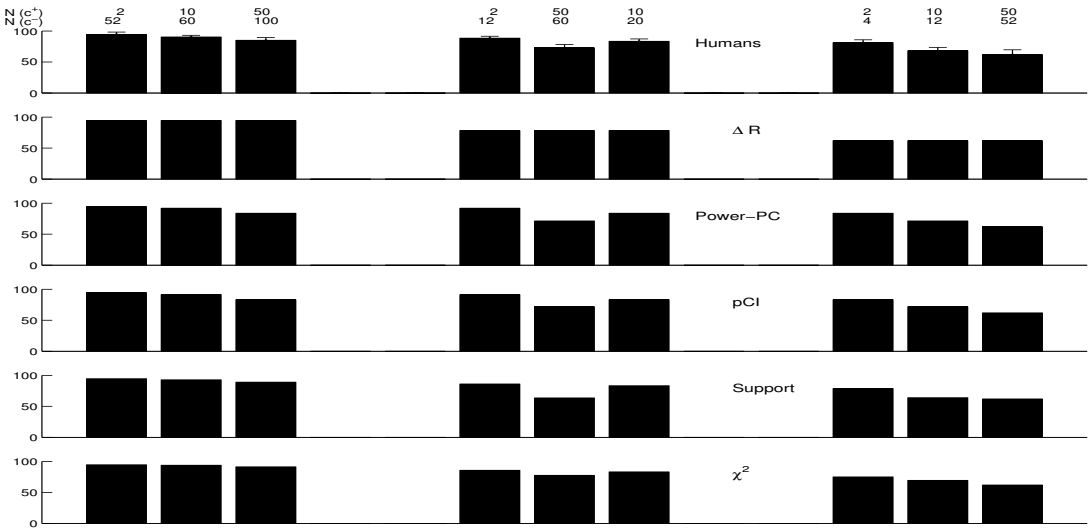


Figure 3.7: Preventative data for particle emissions: human responses and scaled model predictions. Support is the model that results from my framework.

3.9.3 Brief delays and singular, lasting effects

Semmelweis' insight affected the lives of many patients. He changed the state of medical practice persistently and thereby changed the rate at which women died from a preventable disease. But rates were not what inspired his methodological breakthrough.

Rather, while Semmelweis was in Venice on vacation, he heard that a colleague (Jakob Kolletschka) had died after being injured during an autopsy. In particular, shortly after being

$f_{\lambda_0}(\cdot)$.

stabbed in the finger by a scalpel, Kolletschka began developing the characteristic symptoms of puerperal fever. Not having an uterus (let alone an enlarged uterus), conventional medical explanation was at a loss to explain this coincidence, and so it did not bother. But Semmelweis was intrigued — here was the instance of a single person who was exposed a particular instance of a cause (penetrating the skin with a knife covered in "cadaverous matter") and developed the symptoms that his patients were exhibiting. Similar to needing to explain the sickness of the infants by appealing to the same cause, Semmelweis hit upon the notion that the "cadaverous matter" could be the cause. He sought to eliminate that as best he could. Liquid chlorine was his solution, and so he instituted his new method, saving many lives by realising what could be learned from the loss of one.

3.10 Inferring form from tabular data

We have expectations that causes precede their effects with relatively short delays, but in some cases we do not directly observe the data as they occur. There would be no way to distinguish if one had only aggregated rate data over a single time period a short delay from a long delay from equally spaced intervals. The temporal information is lost.

Though exact data often is not observed directly that does not mean we cannot have observations with temporal information. If the event in question can only occur once in an entity and has a persistent state once that event occurs (for example, things die only once and once dead remain dead), rough temporal information can be gained through a sequence of point observations from which you infer what occurred during the time between observation. That is – in the case of entity death – at particular times you can observe the same entity, checking at each time point whether or not the entity has died. Because you know that the death event occurs once, and once it occurs it cannot occur again, the event must have occurred in the interim period. Conversely, if not dead, the entity has not experienced the event in question.

Then, causal inference can proceed in a manner like what I have discussed before, but instead wait times until death after a point of intervention rather than using rates of events under stative interval intervention. Consider that Semmelweis did not observe the death of his colleague, but only heard of it while he was on vacation. The inference that he made to be exactly of this type. But it was an inference made possible only because of the accidental intervention (Kolletschka being stabbed with a recently used scalpel) and the coöcurrence of symptoms characteristic of puerperal fever that preceded his death [⬡].

[⬡] "I could see clearly that the disease from which Kolletschka died was identical to that from which

In cases where such rich background theories are absent, one data point would not suffice. Even if the background theory were rich, if that theory included a high base-rate of death and symptoms, one data-point might not suffice. If the base-rate of death is substantial enough, because death can occur only once, observing a single entity may not give sufficient warrant for any inferences. Semmelweis' inference worked only due to the rarity of the symptoms in the underlying population. In that situation a collection of cases would need to be studied in the same time periods. Instead of recording whether one entity died between two observations, one would record many. In aggregate the wait-time evidence from many samples would be enough to distinguish between effective causes and effects. This is the kind of evidence that was needed for Louis Pasteur to convince people that he had discovered a vaccine for anthrax.

3.10.1 Pasteur and anthrax

Pasteur had already made a name for himself based on his refutation of spontaneous generation and the concordant improvement in brewing techniques that resulted from his realisation that microbes were arriving on dust and growing in the wortDolan [132]. He was attempting to continue his rise to fame by making headway on what he saw to be a related problem — again microbes infecting places that they should not, in this case the problem of disease. He wanted to come up with a systematic way to develop vaccinations.

He happened to get lucky that one of his assistants and he left for vacation at roughly the same time when he learned of the possibility of creating a vaccination against chicken cholera, which accelerated his research programme. His next project was to address the problem of anthrax in sheep.

However Joseph-Henri Toussaint a veterinarian was working on the same topic [191]. He was the first to actually make a chicken cholera vaccine, and he offered samples of it to Pasteur whom he admired. Pasteur did not publicly acknowledge this intellectual debt and when on July 12, 1880, Toussaint revealed success on vaccinating both dogs and sheep from anthrax. Pasteur would not have his thunder stolen; he would put on a spectacle.

Around ten months later, on May 5, 1981 Pasteur and Chamberland [5] arrived at Pouilley-le-Fort to demonstrate their vaccination on the sheep. He had established his method and his hypothesis before hand with an agreement with the Agriculture Society of Melun who would

so many hundred maternity patients had also died.…I was forced to admit that if his disease was identical with the disease that killed so many maternity patients, then it must have originated from the same cause."(from Semmelweis [188]).

provide his subjects. * Even if he did it for scurrilous reasons, Pasteur was admirable in that he make his claims clear and bold.

The agreement was as follows:

1. There were to be 60 sheep, 10 of whom would be left untreated.
2. At an interval of 12 to 15 days 25 of the remaining 50 sheep were to receive two instances of vaccination.
3. After another interval of 12 to 15 days, the 25 vaccinated sheep and the 25 remaining sheep undergoing treatment were to be inoculated with a virulent strain of anthrax.**
4. The sheep would live together in a cattle shed for the remainder of the experiment; the vaccinated sheep would be identified by having a hole punched in their ear.
5. Every sheep that died of anthrax would be buried in separate pits neighboring each other.
6. 25 new sheep would be brought near to the burial site to show that the anthrax is still present and virulent.
7. Another 25 new sheep would be kept near to the location of the study, but far enough away that it can be demonstrated that anthrax was not endemic to the area.
8. Additionally, at the end of the period of observation the vaccinated sheep would be compared to the 10 remaining sheep to ensure vaccination did not prevent the sheep from returning to normal.

As is always the case the actual experiment did not go exactly as planned.

On May 5, the experiments began in front of what would be a regular crowd. Instead there were only 58 sheep, with 2 sheep having been replaced with goats, and 10 cattle (8 cows, 1 bull, 1 ox) included as part of the total sample. That day They inoculated 24 of the sheep, 1 goat and 6 cows with the attenuated anthrax. Twelve days later(May 17) they were vaccinated again, this time with a more virulent (but still attenuated) strain. Two weeks after that (May 31) the whole

* This might be one of the earliest cases of preregistered hypothesis and methods to have occurred in public.

** They added the claim that all the 25 sheep would perish and that not one of the vaccinated sheep would be.

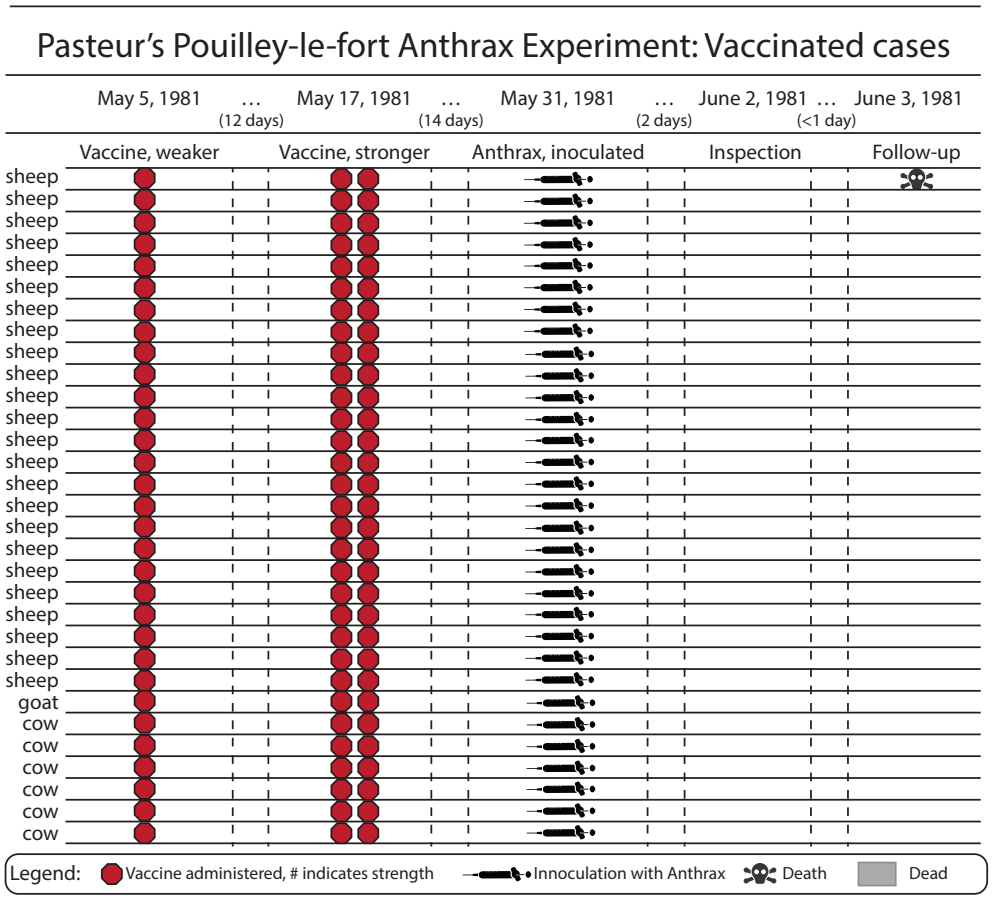


Figure 3.8: Results from the vaccinated condition in Pasteur and Chamberland's[5] anthrax experiments represented in tabular form. The point cause preventative vaccination occurrences occurred in two forms, one stronger than the other and are represented by the red octagons. The innoculation event where the death-generating cause (anthrax) was injected into the animals is represented by a syringe. Death events are represented by skulls and crossbones. And once an animal is dead, the remainder of the days associated with that animal are coloured with grey to indicate their having died, in the vein of Greville and Buehner[2].

set of experimental subjects were injected with fully virulent anthrax from a batch that had lived in Pasteur's laboratory since March 21, 1877.

48 hours later, the crowd returned and every one of the vaccinated animals appeared healthy. 21 of the 24 sheep and the goat died of anthrax in the intermediate days. Were that not enough evidence, two more of the unvaccinated sheep died as the crowd watched. The last sheep died

later that day. The cows did not die (unvaccinated or vaccinated), though only the unvaccinated were visibly ill in that they had liquid-filled edemas growing from where they were injected.

The next day, one of the vaccinated ewes died, and it was discovered that she was pregnant and her lamb died roughly 12 to 15 days prior. Pasteur dismisses this as a death caused by the death of her lamb but does not acknowledge the temporal coincidence of the death of the lamb and the application of the stronger strain of anthrax.

Regardless, by the standards of those who had witnessed it and by the standards of most people, Pasteur had succeeded. He had developed a cure for anthrax. At that point it mattered little that Toussaint had done so almost a year prior.

The form of inference Pasteur, you, I, and everyone who was present at the display engaged in is a powerful form of causal inference. With the exception of the two sheep that died while they were watching, no one knew exactly when the events occurred, but they did know the period during which they occurred. Being alive today, we cannot have observed it, and so we work off of Pasteur and Chamberland's [5] written report of the results like that which can be displayed in tabular form (as in Figure 3.8). And yet we are capable of making the same causal induction: Pasteur had developed a cure for anthrax.

3.10.2 Using decay distributions to model one-shot events

What would making a model that is capable of capturing the kinds of inferences we made in response to Pasteur's data actually entail? It is certainly the case that the timing at which the events occurred mattered, suggesting that the previous approach of having a constant rate over time will not suffice. Thus, we need a formal mechanism for describing how the influence of a cause changes over time. But we also did not see the actual occurrences, but merely heard them reported, so we will need some way of designing the likelihood so that it can take into account our lack of knowledge about the exact occurrence.

I will now describe how I implement causal induction over tabular information of this sort. I will use the decay distributions described in subsection 3.7.4 and will aggregate the effects of those distributions over blocks of time as defined in Equation 3.9. I then apply the resultant model to modelling the results of Greville and Buehner [2] [▽], which studies causal induction on the basis of tabular data with one-shot events.

[▽] I thank William Greville for providing the stimuli needed for these analyses and Marc Buehner for agreeing to their transmission.

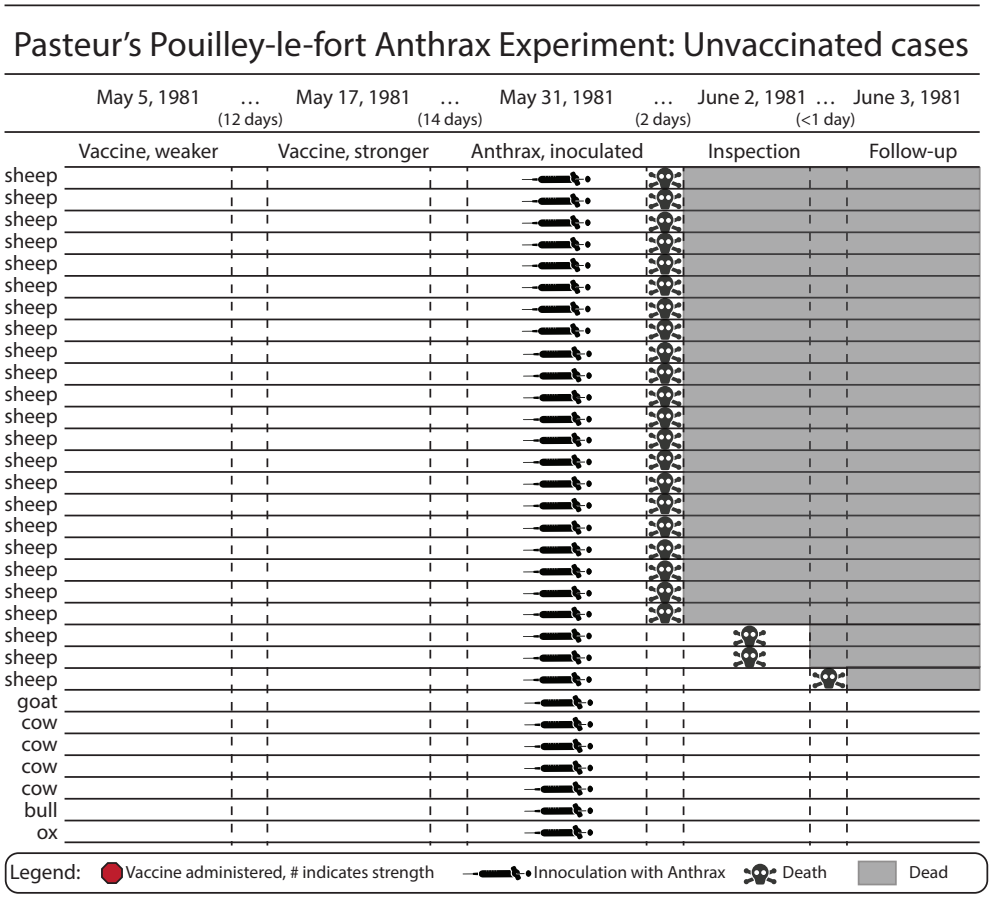


Figure 3.9: Results from the control (unvaccinated) condition in Pasteur and Chamberland's[5] anthrax experiments represented in tabular form. There were no preventative point cause events as this was the control condition. The innoculation event where anthrax was injected into the animals is represented by a syringe. Death events are represented by skulls and crossbones. And once an animal is dead, the remainder of the days associated with that animal are coloured with grey to indicate their having died, in the vein of Greville and Buehner[2].

Tabular data is data like that found in traditional life tables (or in Figure 3.8 which records not the exact time at which events occur but during which of a series of continuous ranges of times the event occurred. This can be seen as one way of recovering trial like structure from continuous-time processes.

I assume that generative and preventative causes are representable by superposition and thinning. I will also assume that decay functions define the degree(or to what proportion) a

cause's influence remains after some amount of time has passed since its occurrence. I will construct this in terms of a base parameter that defines prior scale for the distribution defining the maximum influence of the cause.

Here, I have a simpler version of the Equation 3.12 where only one kind of point cause is needed, all the other work occurs in defining the decay distribution relative to that one event and the manner of aggregating the information across the different samples.

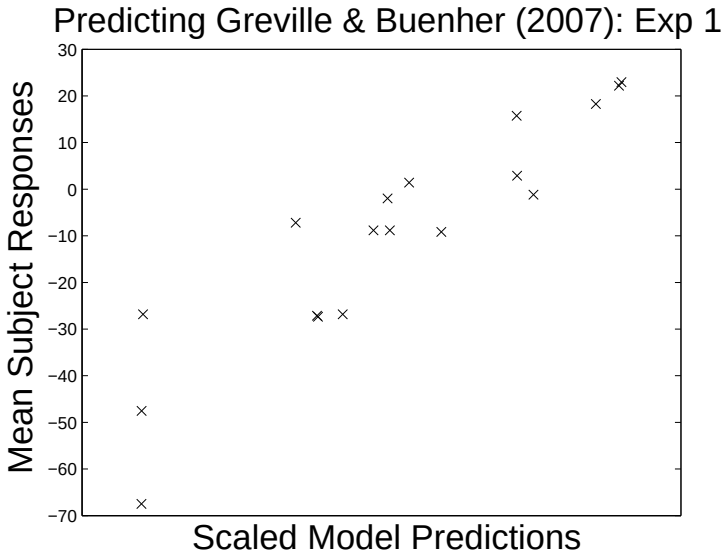


Figure 3.10: Model predictions for Greville and Buehner[2], Experiment 1.

3.10.3 A study in causal induction with tabular temporal data

In each condition of Greville and Buehner's[2] experiments, participants were shown two groups of 40 bacterial cultures, one group which was exposed to radiation and one which was not. Participants were shown (in tabular format) on which of 5 days each batch of bacteria died (if they died). Participants were asked to rate the effect of the radiation on a scale of -100 to 100 where -100 meant that the treatment was very effective at killing the bacteria, while 100 meant that the treatment was very effective at preventing the bacteria from dying (a rating of 0 meant that the treatment had no effect). Though this phrases the problem as one of strength in either direction, I will be modelling it as one of certainty that the graph takes on a particular

form (generative or preventative). This is akin to the distinction discussed at length in Griffiths and Tenenbaum [30].

Greville and Buehner [2] asked each participant about 18 pairs of tables, which differed in the frequency and distribution of times of death. In particular, Greville and Buehner [2] varied the number of cultures dead by day five and the distribution over the times at which the bacteria died. They first fixed the number of deaths that would occur in each table. In all conditions, the time distribution for the bacteria not exposed to radiation was such that each of the deaths occurred with equal probability in any of the five days. However, for the bacteria exposed to radiation there were three time-of-death distributions: "strong contiguity", in which bacteria death was more likely in the first few days after the radiation treatment; "weak contiguity", in which bacteria died more often later in five day period; and "random", in which bacteria death was uniformly distributed among the five days. Contingency information was held constant while varying contiguity. The results of the experiments showed that temporal information dramatically affects human causal inference.

3.10.4 Modelling Greville and Buehner

Modelling the studies in Greville and Buehner [2] requires that I specify the likelihood for 2 conditions of 40 particular one-shot occurrences (bacteria deaths) that share structure within the condition. As such, I only consider a bacterium to have died on the first arrival in the Poisson process defining the rate of death (i.e., $p(t_1 \leq t) = \lambda(t)e^{-\Lambda([0,t))}$, where $\Lambda([a,b)) = \int_a^b \lambda(s)ds$). But, we do not know the precise time at which the bacterium died, merely the day on which it died. Thus I cannot use a form like that in Equation 3.11, but rather must turn to a form like Equation 3.9. Therefore, the likelihood that bacterium i died on day k (that is, the time period $\tau_k = [t_{k-1}, t_k)$) is $\int_{t_{k-1}}^{t_k} \lambda_i(s)e^{-\Lambda_i([0,t))}ds = e^{-\Lambda_i\left([0,t_{k-1})\right)} - e^{-\Lambda_i([0,t_k))}$.

This notation allows λ_i to be different for each culture i, however in this case I will have only two kinds of $\lambda(\cdot)$ functions that will be presumed to be identical within each condition. I am not modelling any properties of each bacterium i as varying in any way. Thus, we will want to determine the likelihood for each i once we know that it either did not die within the observation window or died within day $\tau_k(i) = [t_{k-1}(i), t_k(i))$.

To model the 80 bacterial cultures in each condition (40 experimental bacterial cultures and 40 control bacterial cultures) as 80 first arrivals on independent Poisson processes defined by the culture's condition identity (i.e., control versus experimental) and the underlying graph.

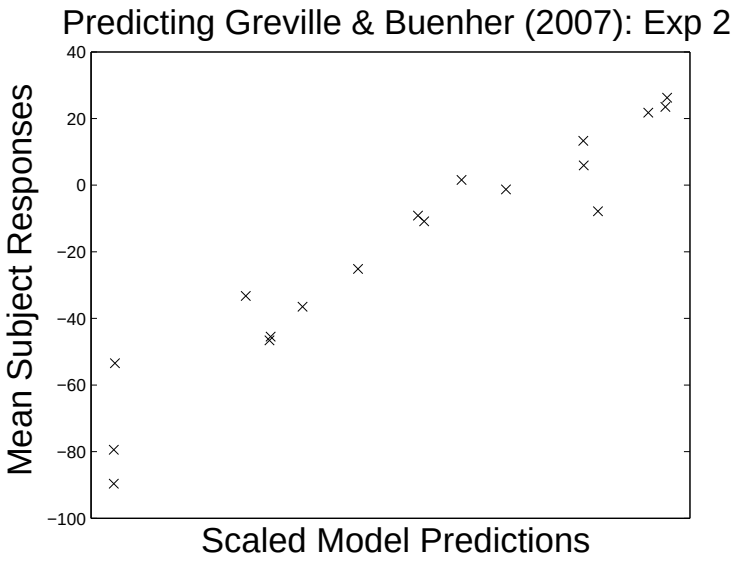


Figure 3.11: Model predictions for Greville and Buehner[2], Experiment 2.

$$p(D|G) = \prod_{i=1}^{80} \int_{t_{k-1}(i)}^{t_k(i)} \lambda_i(s) e^{-\Lambda([0,s))} ds.$$

I obtained the stimuli data for most conditions directly from Greville and Buehner[2♯], but the "random" conditions in their paper did not include exact stimuli. For these cases, I sampled the time of deaths of the bacteria uniformly at random (which was how they had generated their death times) 20 times and took the average of the model predictions across these samples.

Greville and Buehner[2] asked participants to respond on a scale of -100 (the radiation definitely causes death) to 0 (the radiation has no effect) to 100 (the radiation definitely prevents death). This means, we have effectively three graphs that we presume people are reasoning about: G_g, the generative graph (where $\nexists_1 = 0$ and $\psi_1 \in R^+$); G_p, the preventative graph (where $\psi_1 = 0$ and $\nexists_1 \in [0, 1]$); and G_0 the null graph (where $\nexists_1 = \lambda_1 = 0$).

However, we only have judgements on a single scale so we will need to reduce our posterior from the 2 dimensional simplex object down to a single dimension. As in Griffiths and Tenenbaum[1], I modelled participants' mean responses in each condition as $P(G_p|D) - P(G_g|D)$, assuming all three graphs are a priori equally likely.

I assumed a scaled exponential decay function with parameter ϕ was used for both genera-

♯ For which I am grateful.

tive and preventative causes (i.e., $f_i(t, t'; \gamma_i) = f_j(t, t'; \theta_j) = e^{-\phi_1(t-t')}$ where t' is the time that a cause occurs). Because the radiation is only applied to the bacteria once at the beginning of the five days, for G_g and G_p, the only occurrence of the cause is instantaneous and appears at $t = 0$. Thus, for the generative graph,

$$\lambda_{0,t} = \int_0^t \lambda_\varnothing + \lambda_1 f_1(s, 0; \gamma_1) ds = t\lambda_\varnothing + \frac{\lambda_1}{\phi_1}(1 - e^{-t\phi_1}),$$

and for the preventative graph,

$$\lambda_{0,t} = t\lambda_\varnothing(1 - \frac{\eth_1}{\phi_1}(1 - e^{-t\phi_1}))$$

Similar to before, as a prior for $\lambda_\varnothing$ I used $v_0 \sim U(\log(10^{-1}), \log(10^1))$ and set $\lambda_\varnothing = e^{v_0}$. This is a range more reasonable for this problem — rates much larger or smaller than this would be difficult to detect in a dataset of only 80 possible bacteria deaths (though making the range wider does not substantially hurt the model's performance). The remaining priors were defined as $\eth_1 \sim U(0, 1)$, $\lambda_1 \sim \Gamma(1, \lambda_\varnothing)$, and $\phi_1 \sim \Gamma(1, \lambda_\varnothing)$, where the priors are defined in terms of $\lambda_\varnothing$ such that they inherit the scale defined by $\lambda_\varnothing$.

3.10.5 Results and Discussion

Using Monte Carlo integration, I calculated my model's judgements $P(G_p|D) - P(G_g|D)$ for the data in Greville and Buehner [2]. Because the experiments used slightly different methods, I evaluated my model predictions separately for each experiment but concurrently for all 18 conditions within each experiment. My model has a scaled correlation of $r = .910$ ($\alpha = 2.74$) with mean participant responses in Experiment 1 and a scaled correlation of $r = .957$ ($\alpha = 1.72$) with mean participant responses in Experiment 2. These results are solid, but they do not outperform Buehner's [158] used to analyse the same data and had an excellent linear fit for the two experiments, $r = .97$ and $.96$.

The Buehner [158] model is an explicit generalisation of the Power-PC approach to modelling causal inference with tabular data. It required the same kind of extension that was needed to extend Power-PC for rate data, whereby we specify the number of "potential events" that are to be counted. But the rate data had the issue where the number of potential events was arbitrarily defined. The tabular data, on the other hand, has an interpretation for the number of potential events (per bacteria) that can be made specific once given a way of interpreting their scenario.

By looking at how this metric is defined, it can reveal the degree to which the model relies on certain properties of the data in order to make its predictions. This, in turn, allows designing experiments that cannot be explicitly analysed in this way.

The first part that specifies non-arbitrariness is the introduction of an observational unit (1 day) with the total observational period being composed of a finite number of those observational units (5 days). This means that there are 6 potential results for any bacteria, it either never dies, or it dies on one of those 5 days. This gives a specific number of potential outcomes, but not the metric needed to say a specific number of potential *events*. Requires specifying a mapping from the number of potential outcomes to its equivalent number of potential events. Because once a bacteria dies it continues to be dead for the remainder of the observation period, one can define the number of events as being the number of days that it is in the state "dead". In a sense, this is equivalent to (in the rate case) saying that (given a partitioning of the observation period), once an event is detected in a sample, one detects another event in every following time unit for that sample. Then the number of potential events per sample is just the number of days during which something could be dead, and the number of event occurrences is the number of days that it is dead.

Thus, the "nearness" in time of the event to the application of the cause is defined only in terms of the proportion of time within the observation period during which a sample was observed to be dead. That is, the criterion is actually a measure not of proximity to the cause time, but distance from the final observation time. This has the unfortunate feature that as the observation period is lengthened (and all the bacteria die), the ability for this method to detect a difference between the experimental and control conditions disappears. That feature is relatively unimportant in and of itself for understanding causal induction (infinite observation periods are not possible in a finite lifetime), but it does illustrate a problem with defining the distance between the cause and effect times indirectly.

This also highlights a potential problem for the method in dealing with multiple events occurring in the same sample (i.e., if the bacteria could die twice). Because the analogy with traditional Power-PC relies on the persistence of the effect once it occurs(it being a binary event that accrues effect days after occurrence), it is not clear how one addresses events that do not persist but can occur at discrete time-step t though they do not occur at timestep s, where $s > t$. The problem is soluble but it is not clear how one can solve it in a non-arbitrary fashion as was done in Buehner [158].

More concerning for a nonambiguous Power-PC approach is the case where the cause occurs multiple times. Suppose one were to apply radiation more than once to the bacteria part-

way through the observation period. Does this create a whole other set of potential events that need to be included in the consideration? Even for those bacteria that had already died? Does the cause need to be indexed relative to the day of its application — i.e., is the sample space for encoding potential causes necessarily the same space for potential events? If it is so indexed, does that mean the cause can only be applied at the beginning of a day of observation, because – if it is applied at the end of the day – the majority of its *actual* effects may occur on the next day?

3.10.6 Counting sheep and days gone by

What then could a Δ-P and Power pc say about Pasteur's sheep experiment? In it there were multiple inoculations and a period of observation that was nearly 10 times as long as the period it took for all those sheep who reported to have died to die. The causes were of unequal strengths. The periods of time between observations were unequal, making it hard to identify a "natural unit" that could describe the data across all the cases. Is it number of observation periods during which they died? In that case these methods are likely to succeed at finding a difference between them. If it is calculated in terms of the number of days, though, the inference would have been far weaker than our intuitions lead us to believe it should be. If one included the cows as well – not having any means of distinguishing them from the sheep, unlike CTCTS which have a built in ontology for representing just such distinctions – the approaches would support an even weaker inference.

A key feature of Pasteur's [129] method was the multiple applications of the vaccine. This practice was necessary in their perspective due to their theories about why vaccines worked (which turned out to be wrong). The theory was based on an extension of Pasteur's work on fermentation. It stated that there were a trace amount of nutrients that any particular kind of microbe would consume and that its rapid consumption and the consequent replication of the microbes throughout the body in some way caused the disease. Vaccines worked because weakened/weaker microbes would deplete the organism of relevant nutrients for that kind of microbe causing that kind of disease, such that when the later, stronger version of the microbes arrived there would be no nutrients available on which for them to feast.

It was that theory that gave them such great success [♭] even though it had no actual biological

[♭] It is worth noting that Pasteur did not claim to merely delay the arrival of anthrax in the sheep, but to prevent it entirely. In the preventative causes I describe here, all we did was delay the inevitable. True complete prevention will require either a perfect preventer (see minor section 3.7.9.1) or a different notion of prevention that interacts more directly make it such that a causal mechanism is otherwise

basis in reality.

The theory led them to be concerned about the strength of the vaccine and the need to have multiple causal events. They knew that if they made a mistake their patient would die. They did this so as to deprive the disease of nutrients, not to gradually train the immune system on the antigens for ever more virulent disease. This reveals one of the powers of causal theories: so long as they have the right structure and encourage you to attend to the right data you can be completely wrong-headed and still end up doing the right thing.

3.11 Inferring form from continuously streamed data

Real world events are experienced in succession, not in tables. How do people identify the functional form of causal relationships when they occur in real time?

3.11.1 Delayed effects and the Radium Girls

In early 1900s radiation, and radium in particular were touted as miracle medicines that "gave direct energy transfusions to depleted organs" to treat rheumatism, gout, syphillis, anemia, epilepsy, and multiple sclerosis(among others)[††]; it was most prominently advertised as a potential aphrodisiac[192]. As of 1917, this "healthful" substance was also to paint the faces of watch dials for the faint glow it emitted for ease of reading at night. The women who painted these dials for the United States Radium Corporation would lick their brushes to get a finer point, ingesting the radium laden paint as they did so. By 1928 dial painters had been dying from anemia and infections as their jaws rotted away. If they did not die, many found their bones riddled with lesions, dissolving under their own weight. To avoid court, the USRC claimed the women were suffering from were "unfit for more strenuous labour" or "poor health [declining] normally" and cited doctors publicly claiming the symptoms to be the result of "congenital disorders [or] syphilis."[193] Only in 1932 did the American Medical Association remove radium from its "list of positive remedies for internal administration" and only in 1934 did the U.S. Department of Commerce issue any federal guidelines for radium protection.[194]

totally shut down[**] . In that sense it sounds more like Dolan[132] was thinking in terms of frequencies in the vein of the "cover-story" of Lagnado and Sloman[6]. There are echoes of this issue that arise when I address that experiment in section 3.12.

[††] Given that the symptoms of radiation poisoning can resemble rheumatism and syphilis, and that it causes anemia these claims seem particularly misguided.

Today is it well known that these women suffered from radiation poisoning due to their jobs. One of the key factors limiting this realisation was the long delay between the cause of the disease and the onset of symptoms. Several years had elapsed before many of the women began showing symptoms. This confounded the diagnosis of the illness; many died without knowing their jobs were the cause. The delay also hampered (for those women who were still alive) the dial-painters legal case; the statue of limitations for occupational diseases in New Jersey at the time was two years, making nearly every plaintiff ineligible for legal recourse without special permission. The expectation that

The tragedy here was not merely a result of radioactivity, but the inferential problem itself. Until the symptoms manifested, people would have been accused of being paranoid for suggesting – without evidence – that any particular substance would kill them. Inference and explanation in the clinic and courtroom were occurring as the events in question continued to take place. Once it was clear that the dial painting practices was a problem, protective measures were established.

But still, the bias to expect short delays that is intrinsic to the human mind played a role in delaying this process of ensuring safety and justice. How can we formally express that prior and how can we demonstrate its effects in a controlled experimental study?

3.11.2 Our methodological approach

In a standard research program, our primary goal would be to affirm my framework by finding the minimal case that differentiates between two. (e.g., trials with multiple cause instances), and show that my model performs better than it. This is especially the case since – in every analysis so far – my model has at best matched the other models' performance, which, admittedly, already had fairly good fits. But even if trajectory of science looks like a succession of models that outperform each other, that is not the goal.

Our goal is not merely to demonstrate which of two models is a better account of human cognition, but rather to construct a framework that is capable of at least representing the complexity of the real-world scenarios. That those scenarios contain data that escape the representational capacities of earlier models is a pleasant side effect but hardly our goal.

Let us consider changing the bacteria death scenario to see how it can generalise to more general real-world scenarios. In particular, I alter the formal structure of the scenario in four major ways:

- The stimuli were observed in real-time.

- Causes and effects can occur more than once.
- There is no "control condition", participants must infer the base-rate from the same data that supports inferring causes.
- The response scale is made on a 3-simplex rather than a linear scale.

3.11.2.1 Real-time display

Though hardly novel in cognitive science experimentation, real-time displays of data (e.g., video) have not been prominent in previous work on human causal induction from temporally distributed event occurrence data [‡‡], at least not in contrast to the literature relying on contingency tables and counts of events[80,196,35,79,123,197,198]. Even experiments using real objects as stimuli depend heavily on a "trial" manner of dividing up experience[125,142,144,151,145,143,199,200,153]. This is changing and some notable exceptions to this historical trend include Lagnado and Sloman[6], Lagnado and Speekenbrink[7], Bramley et al.[201].

3.11.2.2 Number of events per entity

First, we can look at the number of times various events can occur. As mentioned, there is nothing even in theory preventing the application of a cause multiple times in the original bacteria death scenario. If we were to merely wish to extend beyond previous work, we could ask people to study bacteria death tables that involve multiple applications of the cause. But we can do more.

Death is a "one-shot" event, that once it occurs then it cannot occur again. Consequently there are many bacteria in the tables, but each bacterium can only die at most once. On the contrary, the premise of most point process models (in contrast to closely related *hazard* or *survival-analysis* models) is that we expect to see events occurring many times not exactly once or not at all (though this too can be modelled as a particular kind of point process model). Death is an inappropriate semantic frame for our stimuli if we wish to have multiple events per entity. However we have already seen that radiation can be used to have point events that nonetheless occur multiple times. Points of stochastic luminescence (analogous to that found in fireflies) can play the same role as points of radiation that were counted in the radiation

[‡‡] Real-time displays are crucial tools in the study of Michottean launching experiments and their extensions, but these are primarily spatio-temporal in structure and often applies more to entity-to-entity mark passing accounts of causal relations, which alters the situation considerably[165,195].

study. However, they allo considering delay distributions, because – unlike summed counts – we can display the actual (simulated) times of occurrence in real-time.

3.11.2.3 Continuous data flow

In most real cases, events are not well partitioned into "control" and "experimental" conditions following standard statistical advice about experiment design to reveal causation. However, the data from which we infer causal relations occurs continuously without any way to separate into trials, let alone the strictures expected by the recommendations of experimental design theory. This situation deserves careful consideration as it reveals a great deal of the implicit theorising that has taken place in standard scientific practice long before any data was collected. Given that the gold standard for determining which samples will be drawn from the experimental and control conditions is randomisation, this bakes in the assumption that that method for performing the randomisation is valid. The randomisation mechanism may be entirely reasonable (e.g., determined by a randomly generated number), but it points to the goal of the entire procedure: to effectively intervene on the system and render the data derived under the experimental condition independent of the data derived under the control condition, apart from those aspects of the manipulation that the experimental and control conditions are designed to distinguish.

However, the control usually acts as a "baseline" case against which the experiment is going to be compared. This is the basis of null hypothesis testing, where the control is the "null" hypothesis. But if you have an explicit characterisation of the "null"/"baseline" case and you can formulate your hypotheses in terms of how the cause will alter the baseline case and how that influence changes over time, you can reason precisely even without an explicit distinction between control and experimental condition. I.e., if you presume that the cause only has a substantial effect for some finite period of time, once substantially far away from the cause's occurrence you can effectively treat the data generated as if it were from the baseline (relative to the specific cause under consideration).

This perturbation from baseline account of experimental versus control conditions is a key insight as to the the utility of the noisy-OR, noisy-AND-NOT, superposition and thinning approaches to defining causal relations. This is in contrast to treating the problem as one in which you are identifying generic *some* probabilistic dependence[78], though people can also reason in that manner[1]. This general notion is different from the particular prior beliefs people have about the parametric form of generative and causal relations[202]. This form of prior infor-

mation presumes the existence of a baseline with causes perturbing the baseline and appears in work on causal attribution such as the "abnormal conditions focus" model of causal attribution[203] and the "Rational Scientist" account of the side-effect effect[204]. Such a formulation may in some case be necessary to be able to account for why (given that the world consists of a uninterrupted continuous stream of events) we are capable of inferring that a cause has *ever* ceased having an effect.

Furthermore, it may not be that all perturbation-type causal functional forms can be reasoned with equally well, as the failure of participants in Experiment 1 of Lagnado and Speekenbrink[7] to identify a "hastening" relationship shows. "Hastening" can only be defined relative to a base-rate, but nonetheless at least in 1-shot causal relations, there is not enough information for people to have identified a "hastening" cause's influence.

3.11.2.4 Response values: posterior probability between three models

Rather than compressing human responses to values along a single scale, I can ask participants to engage in a task more directly analogous to my model: assigning a probability distribution to a set of graphs. In some cases, this may consist of choosing whether a causal relation exists as well as its form or it may consist of choosing among various models. Alternatively, as above, you could integrate information across the posterior based on some function of the graph, which allows you to reason simultaneously about the larger causal system (as expressed in the posterior over graphs) and elemental causal relations (the summarised belief when considering the graphs containing a particular edge).

Regardless, providing people the opportunity to express their beliefs in terms of probability distributions allows more directly modelling their answers in the terms of probabilistic inference. In particular, I can directly compare their responses to my model's predictions rather than merely associating them on a linear scale.

3.11.3 Experiment design: Continuous time bacteria

The experiments that follow can be seen as an extension of the scenarios I have discussed before in which I try to address some of the structural problems with the scenarios and their data and ask people to engage in a more realistic exercise of inference. I have described a variety of properties that are formally useful; it would be useful to build a model of an experiment that relies on those properties. Now, I wish to extend the bacteria death scenario described in Gre-

ville and Buehner [2] to an experimental scenario outside the scope of the causal power for rates model [158].

3.11.3.1 Stimuli.

The stimuli I created in other ways were very similar to those in Greville and Buehner [2]. I presented each participant with 18 within-subjects conditions each of which had a cause indicator and 40 bacterial cells which I told participants could not affect one another. Participants were told that the bacteria light-up spontaneously, but that scientists wanted to know whether different kinds of radiation changed the rate at which they lit up. I generated the data for the stimuli using a Non-homogeneous Poisson process defined as in Equation 3.12, with $\lambda_0 = \frac{2}{5}$and $\phi = \frac{1}{4}$ in all conditions. The 18 conditions were defined by the cross between number of cause occurrences $\{1, 2, 3\}$, by type of true causal relationship {generative, preventative} and the strength of the causes {weak, medium, strong} (particular parameter values for each strange can be found in Table 3.2).

Participants were told that they were to imagine that they were assisting a biotech lab that is testing radiation treatments on bioluminescent bacteria. That they would be watching a series of videos in order to

> … examine this data and make a determination about whether the radiation increases the rate of bioluminescence, decreases the rate of bioluminescence or does not affect the rate of bioluminescence.

Each video lasted thirty seconds and had one potential cause of 40 luminescent bacteria cells. If the cause occurred once, it occurred at $\{15\}$ seconds; if it occurred twice it occurred at $\{10, 19\}$ seconds and if it occurred three times it occurred at $\{3, 8, 20\}$ seconds. Thus, participants observed the forty bacterial cells and one cause cell lighting-up within the duration of thirty seconds in each condition for each of the 18 conditions. They had an opportunity to rest between each condition. They were told that each condition was independent of the others. The order of the conditions was randomly assigned. The display refreshed at a rate of 50 fps, though because of limitations in timing events running in the browser, I need to empirically estimate how closely the display matched these expectations.

Because we cannot produce, let alone perceive, instantaneous events, we need to ensure that every event occurs for some amount of time. For the purposes of my model, I will still treat these as point events that occur at the onset of each event. Because of the relative importance

of observing the single cause node, each cause lasted 100 ms while the events occurring on the 40 effect channels each lasted 80 ms.

Parameter Values					
Generative, ψ_i			Preventative, $\nexists_j$		
Weak	Moderate	Strong	Weak	Moderate	Strong
$\frac{1}{4}$	$\frac{7}{20}$	$\frac{3}{5}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$

Table 3.2: Values for the causal parameters used in generating real-time stimuli for section 3.13

Challenges for programming dynamic stimuli in ECMAscript. It is worth noting that web browsers generally will not provide accurate ECMAscript based timers at all. This problem is made worse when animating events at 50 fps. In particular, if the browser is under heavy computational or memory load, the browser's internal clock will gradually drift out of sync with the computer's internal clock which I take to be veridical. But the browser will display events according to its internal clock, as that is what ECMAscript timed events are linked to. Fortunately, because modern browsers also allow access to the underlying computing system's clock, it is possible to identify how far the browser's clock has drifted on each "tick" of the internal clock. This allows not only knowing when you are out of sync but because you can reset the internal clock's value, you can modify it to be in sync with the system clock, recallibrating it on the fly. Furthermore, I can store this information about the amount of recalibration process for each participant, allowing me to ensure that I only include data from participants who had approximately accurate depictions of event series. Though these differences are (usually) on the order of milliseconds, which may seem minor, with so many events occurring these lags can rapidly add up.

I will refer to the amount of recalibration needed for each event in terms of the "lag" for that event. Then, for each participant per condition I can average the "lag" recorded for each event in the condition (which is just the arithmetic mean $\frac{1}{K}\sum_1^K$). This provides a set of "mean-lags" for each participant indexed by condition the participant was in when experiencing each "mean-lag". From this I compute a per participant average "mean-lag" and a max "mean-lag". I will use these values to exclude participants if they exceed an error threshold for either the average or max mean-lags. Based on pilot testing, I established a 5ms threshold for the average mean-lag and a 100ms threshold for the max mean-lag.

3.11.4 Response measures: eliciting probability distributions

Participants were told they would be rating their confidence in causes having a certain form by distributing 100 points and were trained on an example of providing a probability distribution over a die roll:

> After viewing the experiment, you will see a rating form similar to the one below. Following the experiment, the form will be titled "Ratings by Type of Relationship" instead, and the fields will be "Causal," "Preventative", and "None". But for the sake of example, pretend you roll a die. Distribute a total of 100 points among the inputs in accordance with your confidence in each outcome and press submit.
>
> __ Rolling 1, 2, or 3 __ Rolling 4 or 5 __ Rolling 6
>
> Note that a good response for confidence in getting 1, 2, or 3 might be "50". When you feel you understand the rating system, click "Proceed to the Experiment". Please make sure you understand the instructions before proceeding, as you will not be allowed to return.

If participants tried to move forward without having correctly assigned 100 points, they were not allowed to do so.

I explicitly did not ask participants to predict what the die roll would be but rather to rate their confidence that it would fall into each category. This type of confidence rating task seems to have the character where the computational level response is to probability match in that you are making a judgement about a probability distribution that will generate data rather than about the data itself.

They then observed the video for a condition and were asked to do the same confidence distribution task but now with regards to the functional form of the causal relationship between the potential cause and the data they had just observed:

> Please distribute 100 points among the categories below[Causal, Preventative, None] to indicate how you consider the radiation to have affected luminescence. Remember a 0 means you have no confidence in a particular relationship and 100 means you have absolute confidence in said relationship.

3.11.5 Modelling and Inference

My model will treat this as a problem of functional form induction over three potential graphs generative(G_+), preventative ($G_\%$)[§§], and null ($G_\varnothing$) graphs. These will be as in the previous two models, where the generative relationship is treated as a superposition of processes initiated by point events, and the preventative relationship is treated as thinning processes initiated by point events, and the null graph says there is no causal relationship or edge between the nodes, explaining all data in terms of an underlying base-rate.

The only processes that can produce events are the base-rate process governed by $\lambda_\varnothing$ and processes initiated when a generative cause occurs (i.e., relative to G_+). The only processes that can cancel events are those initiated by preventative cause occurrences. The difference between generative and preventative functional forms (particularly, their additive versus multiplicative characters) alter the way influences from multiple causal instances needs to be taken into account in the likelihood. The generative causes are superpositive/additive (as discussed previously), which means that the influence of multiple causes is easily decomposable. The preventative causes are multiplicative, and as a result they cannot be decomposed as easily and we need a slightly more complicated functional form when dealing with them.

In order to compute the likelihoods under multiple cause events it is convenient to split up the data set into those events that occur before, after and between cause events. Because the influence of a cause cannot project backwards in time, this allows us to divide the system up into smaller pieces that will be governed by the same causal laws since (by definition and hypothesis) no other causes occurred within the span of time.

Notation: ς. Because we know that at most there is one causal process and many potential events, I will distinguish the times at which causes occur and the times at which events occur. Stigmas (ςs) [¶¶] will represent in this case the cause events whereas ts can represent the times of effect events and the analogous times for the cause events when we need to think in terms of intervals ($\tau = [t_1, t_2)$).

[§§] Note, this is $G_\%$ and not G_- as might be intuitive in order to reiterate that the functional form of prevention is to cancel some percentage of the events that occur, not to decrease the absolute rate of occurrence.

[¶¶] ς is a ligature combining the characters sigma (σ) and tau (τ). I am using it partially because it occurs rarely in mathematics, and partially because its name refers to "a mark" in the original Greek, which makes it appropriate for representing point events.

3.11.5.1 *Kinds of events: what is inferred*

I will be treating the cause events as fundamentally different from the effect events. In particular, I will be performing inference on the effect events taking the cause events as given, or – more accurately – as having occurred because the cause was intervened upon, rendering the cause occurrences as independent of one another and perfectly likely.

3.11.5.2 *Splitting the set of point events*

The likelihood for sequence of fully observed events can be decomposed into the likelihood for the component events and their intervals. For that, we can rely on Equation 3.11, and we can define the loglikelihoods once we have the intensity(λ) and intensity measure (Λ) functions. For further details see minor section 3.7.11.1.

3.11.5.3 *Multiple effect sequences*

Because I will need to consider 40 different effect sequences, I will need to do this event-sequence splitting separately for each of the 40 sequences, and then take the product (sum) of their likelihoods (loglikelihoods).

Worth noting is the fact that the observation period and the causes will be identical for each of the event sequences meaning I will need a separate $\{\mathbb{t}\} \stackrel{\text{def}}{=} \text{sort}(\{t_0, t_{\text{fin}}\} \cup \{\varsigma_1, \dots, \varsigma_n\} \cup \{t_1, \dots, t_k\})$ for each effect process. Fortunately these are going to be treated as being independent conditional on the causes and the observation period, leaving us with

$$\ell(\{\{\mathbb{t}\}_1^{40} | \{\varsigma\}_1^n, \{t_0, t_{\text{fin}}\}, \Theta) = \sum_{i_2=1}^{40} \ell(\{\{\mathbb{t}\}_{i_2} | \{\varsigma\}_1^n, \{t_0, t_{\text{fin}}\}, \Theta). \tag{3.13}$$

3.11.5.4 *Generative-cause Likelihood model*

In my earlier generative causal models, I had either rates or a delays until a single event, whereas here I have multiple effect events. So I will need to analyse the likelihood function for multiple events. Similarly, in the previous sections I only had one cause "event" (section 3.10) or one "period" during which the cause was active (section 3.9), whereas here I can have multiple causal events. Fortunately, the formal process of splitting the event set described in the previous subsubsections will give us a logliklihood form that I can work with.

The generative model is additive and so its intensities and intensity measures are easily decomposable. I will be using exponential decay functions, though that is not necessary.

Specifically, then, if there are k generative causal events preceding t at times $\{\varsigma_1, \varsigma_2, \dots, \varsigma_k\}$. There is an underlying rate of $\lambda_\varnothing$, a max intensity of ψ with an exponential decay governed by ϕ:

$$\lambda(t|\{\varsigma_i\}_1^k, \Theta) = \lambda_\varnothing + \sum_{i=1}^{k} \psi \exp(-\phi(t - \varsigma_i))), \tag{3.14}$$

and the intensity measure would be,

$$\begin{aligned} \Lambda(t_1, t_2|\{\varsigma_i\}_1^k, \Theta) = & \int_{t_1}^{t_2} \lambda_\varnothing + \sum_{\varsigma_i \in \{\varsigma_i\}_1^k} \psi \exp(-\phi(t - \varsigma_i))dt \\ = & \lambda_\varnothing(t_2 - t_1) + \sum_{\varsigma_i \in \{\varsigma_i\}_1^k} -\frac{\psi}{\phi}\left(e^{-\phi(t_2 - \varsigma_i)} - e^{-\phi(t_1 - \varsigma_i)}\right). \end{aligned} \tag{3.15}$$

3.11.5.5 Preventative-cause Likelihood model

We can use the same method of considering the loglikelihood of no event occurring during the time over which no event occurred($[t_1, t_2), \ell(N([t_1, t_2)) = 0|\Theta, \varsigma_{i_1}^k)$) and the likelihood of an event occurring at the time at which an event occurred $t_2, \ell(N(t_2) = 1)|\Theta, \varsigma_{i_1}^k)$.

To define these, I will need to define the total intensity function $\lambda(\cdot)$ and its integral, the total intensity measure $\Lambda(\cdot, \cdot)$.

I will use the same functional form as for earlier prevention, but now that more than one event can occur it requires quite a bit more accounting.

Suppose there were k preventative causal events that preceded t at times $\{\varsigma_1, \varsigma_2, \dots, \varsigma_k\}$. There is an underlying rate of $\lambda_\varnothing$, a max prevention probability of $\eth$ with an exponential decay governed by ϕ. Then,

$$\lambda(t|\{\varsigma_i\}_1^k, \Theta) = \lambda_\varnothing \prod_{i=1}^{k} (1 - \eth \exp(-\phi(t - \varsigma_i))), \tag{3.16}$$

which is a straightforward consequence of the previous functional form. However, the total intensity measure is substantially more complicated because the causes interact with one another due to their multiplicative functional form. Supposing that all the cause events precede

the lower boundary of the interval in question(i.e., $\tau = [t_1, t_2), t_1 < t_2$ and $\{\varsigma_i\}_1^k, \forall \varsigma_i < t_1$):

$$\begin{aligned}\Lambda(t_1, t_2|\{\varsigma_i\}_1^k, \Theta) &= \lambda_\varnothing \int_{t_1}^{t_2} \prod_{\varsigma_i \in \{\varsigma_i\}_1^k} (1 - \eth \exp(-\phi(t - \varsigma_i)))dt \\ &= \lambda_\varnothing \int_0^{t_2 - t_1} \prod_{\varsigma_i \in \{\varsigma_i\}_1^k} (1 - \eth \exp(-\phi(t + t_1 - \varsigma_i)))dt \end{aligned} \tag{3.17}$$

To get to an analytic form of this integral, we need to expand the product over the k terms. This leads to a sum of 2^k terms (1 term for for each set in the powerset of the multipliers)

$$\prod_{i=1}^{k} (1 - \eth e^{-\phi(t+t_1-\varsigma_i))} = \sum_{\mathcal{T} \in \wp(\{\varsigma_i\}_1^k) \setminus \varnothing} \frac{\eth}{|\mathcal{T}|\phi} \left(e^{-|\mathcal{T}|\left(t_2 - \frac{\sum_{u \in \mathcal{T}} u}{|\mathcal{T}|}\right)} - e^{-|\mathcal{T}|\left(t_1 - \frac{\sum_{u \in \mathcal{T}} u}{|\mathcal{T}|}\right)} \right). \tag{3.18}$$

To see why, consider just the case where there are three events, giving us

$$\begin{aligned}\prod_{i=1}^{3}(1 - \eth e^{-\phi(t+t_1-\varsigma_i)}) =& 1 + (-\eth)^1 e^{-\phi(t+t_1-\varsigma_1)} + (-\eth)^1 e^{-\phi(t+t_1-\varsigma_2)} \\ &+(-\eth)^1 e^{-\phi(t+t_1-\varsigma_3)} + (-\eth)^2 e^{-\phi(t+t_1-\varsigma_1+t+t_1-\varsigma_2)} \\ &+(-\eth)^2 e^{-\phi(t+t_1-\varsigma_1+t+t_1-\varsigma_3)} + (-\eth)^2 e^{-\phi(t+t_1-\varsigma_2+t+t_1-\varsigma_3)} \\ &+(-\eth)^3 e^{-\phi(t+t_1-\varsigma_1+t+t_1-\varsigma_2+t+t_1-\varsigma_3)}.\end{aligned} \tag{3.19}$$

Which can be simplified to (with $\mu(\cdot)$ denoting the arithmetic mean),

$$\begin{aligned}\prod_{i=1}^{3}(1 - \eth e^{-\phi(t+t_1-\varsigma_i)}) =& 1 - \eth e^{-\phi(t+t_1-\varsigma_1)} - \eth e^{-\phi(t+t_1-\varsigma_2)} - \eth e^{-\phi(t+t_1-\varsigma_3)} \\ &+\eth^2 e^{-2\phi(t+t_1-\mu(\{\varsigma_1,\varsigma_2\}))} + \eth^2 e^{-2\phi(t+t_1-\mu(\{\varsigma_1,\varsigma_3\}))} \\ &+\eth^2 e^{-2\phi(t+t_1-\mu(\{\varsigma_2,\varsigma_3\}))} - \eth^3 e^{-3\phi(t+t_1-\mu(\{\varsigma_1,\varsigma_2,\varsigma_3\}))}.\end{aligned} \tag{3.20}$$

More generally this function is in multi-binomial form (i.e., in multi-index notation $(x - y)^\alpha = \prod_{i=1}^{k}(x_i + y_i)^{\alpha_i}$), where each element is raised to the first power ($\alpha_\cdot = 1$), the first term in each is $1(x_\cdot = 1)$, and the second term is negative, meaning that the valence

of the sum elements that it outputs will depend on the number of the second terms that are multiplied to produce that sum element($y_i = (-1)e^{-\phi(t+t_1-\varsigma_i)}$). I express the expanded version of the product exactly, which leaves us with a sum of terms that can individually be integrated using the standard form.

The general form is (with $\mathcal{T}$ denoting subsets of the set of causal events $\{\varsigma\}_1^k$and n being the number of events included in each term**, i.e., $n \stackrel{\text{def}}{=} |\mathcal{T}|$),

$$\prod_{i=1}^{k}(1-\nexists e^{-\phi(t+t_1-\varsigma_i)}) = 1 + \sum_{\mathcal{T}\in\wp(\{\varsigma_i\}_1^k)\setminus\varnothing} \frac{(-\nexists)^n}{n\phi}\left(e^{-n\phi(t+t_1-\mu(\mathcal{T}))}\right). \tag{3.21}$$

Which, can then be plugged in to give us the intensity measure for no events occurring in the interval $[t_1, t_2)$ under the influence of $\{\varsigma_i\}$ cause events that all precede t_1. It has the form the form

$$\begin{aligned}\Lambda(t_1, t_2|\{\varsigma_i\}, \Theta) &= \lambda_\varnothing \int_0^{t_2-t_1} \prod_{i=1}^{k}(1-\nexists e^{-\phi(t+t_1-\varsigma_i)})dt \\ &= \lambda_\varnothing(t_2-t_1) + \sum_{\mathcal{T}\in\wp(\{\varsigma_i\}_1^k)\setminus\varnothing} \frac{(-\nexists)^n}{n\phi}\left(e^{-n\phi(t_2-\mu(\mathcal{T}))} - e^{-n\phi(t_1-\mu(\mathcal{T}))}\right).\end{aligned} \tag{3.22}$$

3.11.5.6 Scaling model predictions

Because the model receives all the data, which is occurring faster than humans can perceive in its entirety the model makes much more certain predictions than people. To account for this I use a method like the forgetfulness used in Bramley et al.[205]. I fit my model to people's data for a variety of proportions ranging from 1 to $1/100$ and found $\frac{1}{13}^{\text{th}}$ to be the best fit.

In terms of the figure, the lower the ratio of remembered data, the closer the model predictions move toward the prior distribution over graphs(which generally is close to the midpoint of the graph).

** Note, I had to extract the term that would be equivalent to the emptyset to avoid division by 0.

3.11.5.7 *Sparsity Prior*

Though there isn't much in the way of variation between the different graphs, the null graph does have one less edge than both the generative and preventative graphs. I computed a sparsity prior that favours having fewer edges by having the form $p(G_.) \propto \omega^{(\#(\text{edges}))}(1-\omega)^{(\#_{\max}(\text{edges}))}$. In this case the best fit is provided by $\omega = .4$ (similar to the scale parameter, this was fit to people's data).

In terms of the figure, the lower the sparsity parameter the more $G_\varnothing$ is favoured.

3.11.6 Results

I collected data from 129 individuals, of whom 75 completed any study conditions, and of that 75, 44 completed the entire set of conditions and had mean-lags that met the 5ms and 100ms average- and max-meanlag thresholds that I set before beginning the experiment.

If this were an exclusion criterion this would be a large number of participants who were excluded, but it is actually a statement about whether people participated in *my* study as opposed to *some* study. That is, those where were excluded did not observe the stimuli that were intended and it is not clear from the callibration measures how to determine what the timing of the stimuli that were presented actually were. However, there is a slight concern in that a worse performing computer is likely to be older and cheaper than a higher performing computer, meaning that the criterion may inadvertantly exclude individuals from a lower socio-economic status groups. The same could be said of any web-based experiment as they require access to a computer and the internet to be a participant; this is just an amplified version of that bias since it relies on people having a fast enough computer to keep up with the animations. Additionally, given the variance observed in the mean-performance times it does not seem that computational limitations were in general a cause for concern; it seems that the participants were completing some other task while participating in the experiment.

So, indirectly, I may have discovered a way to detect whether participants are actively doing other tasks while participating in your study (in as much as those other tasks cause delays in the observed stimuli). This may be useful for others who wish to build web experiments that require continued attention during the course of the experiment.

My model has an $R^2_E = .787$, which under standard interpretations suggests that I have explained 78.7% of the variance. Because I am predicting values on a simplex, a Euclidean distance metric is not necessarily appropriate. The model's KL divergence with the human mean

values is 0.0673, whereas the "mean" model's KL divergence is 0.2799.[✠✠] Unfortunately, there is no standard way of associating describing KL divergence in terms of "variation explained".

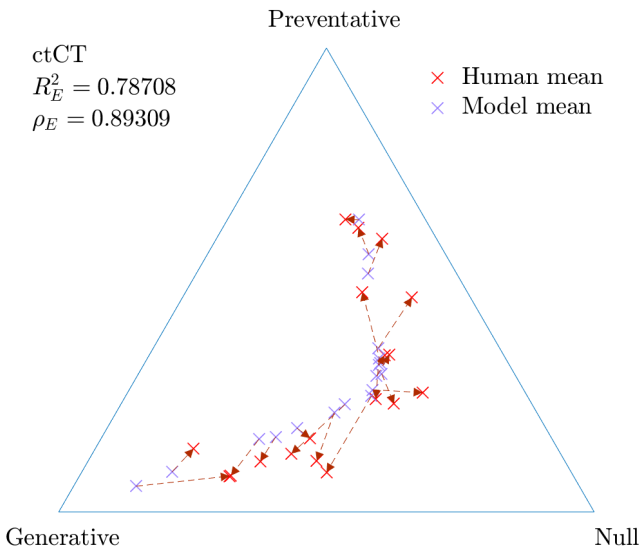


Figure 3.12: Results from Continuous time bacteria experiment.

3.11.7 Repeated causes and repeated effects: radium

Only the women who worked at the United States Radiation Corporation for more than a year between the years 1917 and 1926 succumbed to radiation poisoning. Only after many days while working at USRC did the women absorb enough radium to produce fatal symptoms and death. The manner of death occurred not because they were exposed to radium, per se, but due to the frequency and the method of conveyance. It was well-known even by the practitioners of mild radium therapy that exposure to large amounts of radium for even moderate periods of time could prove fatal[192]. Nonetheless, many thought that imbibing small amounts of radium was beneficial to health (or so were the claims of those peddling radium-laced patent medicines).

Fellow co-workers did not suffer the dial-painters fate. It was not a general ambient exposure to radium that proved fatal, but multiple instances of particular causal events When their

[✠✠] It may be even more appropriate to use a non-Euclidean distance metric such as Aitchison[206],[207] distance, however doing so will require an analysis beyond the scope of that described here.

brushes left their lips after being pointed, the paint and the radium within remained. When consumed orally – especially through exposure to the mucous membranes of the mouth(not the stomach) – radium is absorbed directly into the blood stream and is rapidly deposited in bones, where it remains. , . "Because patients lived three to eight years or more after they left work as dial painters, the radio-active substances at time of death were confined entirely to the bones."[208] Once they had been in their positions for so long across many causal instances, there was little anyone could do for them. They no longer needed to work at the dial painting plant to be exposed, but carried around their exposure with them. "After final deposition in the bones, these deposits emitted their characteristic radiations day after day, month after month, and year after year, diminishing in amount only by their uninfluenceable natural decay."[208]. Their deaths then were not due to single causal events or event single effects. They were the accumulated damage of many small events of alpha decay over years of their bones harbouring the radioactive poison. Fortunately, if a victim did live past six or seven years from initial exposure, the poisonous deposits would have diminished more than half due to natural decay. The effective rate of effect due to each instance the brush was placed in the would have been reduced by roughly 70 percent; sparing the women's lives if not their skeletons.

People are able to reason about these complicated mechanisms because they are capable of reasoning about systems in which multiple instances of causes have multiple instances of effects. We can reason, too, about the decay in the rate of an effect due to these multiple causes. And, fortunately, data of this sort can convince us from thinking something is a prophylactic to the deadly hazard it actually is.

3.12 Inferring structure in hidden mechanisms from trials of one-shot events

> *I consider that the cause of cholera is always cholera; that each case always depends upon a previous one.*
>
> John Snow, 1855[209]

The causal relationships studied in the previous sections are relatively straightforward in the kinds of relationships that we posit and the way that the observations available to reasoners reflect those relations. However, that is not always the case. In fact, not being able to observe the intermediate mechanisms could be said to be *the* hard problem for causal induction, as it

would seem to violate empirical accounts of causal induction since the mechanism by definition is not observed.

This becomes even more difficult when you lose a property that the other analyses rely on: specifically, the kinds of entities that can be causes or effects are obviously distinguished from each other. In experiments like mine no one could ever mistake the application of radiation for the death of a sample bacteria. In the real-world, causal mechanisms are far from so opaque and the causes and effects that determine the course of events often need to be identified on their own. One of the most insistent critiques of John Snow's theory of cholera's contagion by water was his insistence on the notion that disease (and therefore death) would *not* arise directly from the exposure via the air to particles of decaying/dead organic material[209] and that it only arose by virtue of transmission of a particular kind of material[⏀⏀]. No one observed the underlying mechanism in either case, but the mechanism in question was a key part of the argument.

When intermediate nodes are unobserved (and therefore of unknown kind) – or when the same kind of thing can play both role cause or effect – we cannot rely on priority information or a priori knowledge to tell us which things are candidate causes and which are effects (or at least not a priori knowledge in the sense of concrete realizations; framework knowledge will prove to be crucial). More precisely, this shows that we had been relying on a priori information not in order to *populate* the potential structure, but in order to *prune* the set of structures. When there is no distinction between the kind of thing that can be a cause and the kind of thing that can be an effect it just means that many more potential structures need to be considered.

3.12.1 Incubation, healthy carriers and Mary Mallon

Even when the causes and effects are of the same kind of thing, there can be the issue of the cause not expressing a feature despite being able to pass along its effects, and even then, only noisily. For example, Mary Mallon (b.1869 – d.1938) was a cook in the United States better known as Typhoid Mary. At the time, it was known that typhoid fever was spread through the action of a microbe; it was assumed to spread through contaminated water sources in analogy to cholera. At the turn of the century cases of "healthy carriers" (individuals capable of transmitting the disease while exhibiting no symptoms) were discovered; Mary was a healthy

[⏀⏀] And note how wrong he was if we compare his claim regarding "cadaverous material" to what Semmelweis discovered regarding the ætiology of childbed fever.

carrier.

Most often, Mary did not infect those for whom she cooked, because the food was prepared at a high enough temperature to kill the bacteria. Unfortunately, the dish the family was most fond of was her "ice-cream with fresh peaches cut up and frozen in it." This preparation had no heat and so it did not disinfect the food; "no better way could be found for a cook to cleanse her hands of microbes and infect a family."[210]. In this way, she is believed to have infected at least fifty-three people, of whom three died, despite showing no symptoms herself.

A less extreme case echoing asymptomatic carriers are diseases for which their incubation period (the time between the moment of exposure and symptoms are express) and latency period (the time between the moment of exposure and when the individual can transmit the disease) diverge. ⋈⋈ This is the case for Human Immunodeficiency Virus (HIV), where no symptoms can be present (though they may later manifest), and meanwhile the individual is still capable of transmitting HIV.

This type of disparity between the onset of symptoms and onset of infectability is occurs in both biological and informational viruses as well (such as those that affect a computer). One of the earliest computer viruses, Elk Cloner, would increment a count of the number of times the device was booted and only show a message on the fiftieth time it was opened. In the meantime, it would copy itself onto any disk it had access to. Other viruses will actively produce symptoms (in the sense that undesired activity occurs immediately upon infection) while hiding the symptoms and the activity from from the user's perspective. Some viruses actively mask their activity in order to achieve this end ⊛⊛.

The important lesson to take from these cases is that temporal information from the effects of the causal processes may not veridically reflect the information of the underlying causal event series that produced the effect events. This is made further complicated when one does not know *a priori* which variables are even potentially causes of which other variables, especially since those causes may be hidden variables as well. Finally, it may be that the causal relations are not primarily defined in terms of the temporal relations. Casual relations might be defined in terms of the the absolute probability that a message will be passed which cannot be expressed in terms of wait-times since infinite wait-times are assumed to have $p(w = \infty) = 0$.

⋈⋈ Note that some diseases will not be transmissible unless symptoms are showing. The variety of how diseases manifest makes it all the more important that we consider myriad of causal mechanisms and functional forms to ensure that we know the particular kind of disease we are dealing with before making any prescriptions.

⊛⊛ This has even inspired a model of computer virus detection based on models of human abductive inference.[211]

In order to show how my framework can handle these sorts of complications I will model the results from Experiment 1 in Lagnado and Sloman [6] .

3.12.2 Experiment Description

In Lagnado and Sloman [6] they asked participants to intervene on a (graphically displayed) small computer network to figure out which connections between the computers were working.[ʊʊ] Each participant would observe 4 different networks that in actuality were governed by the same structure, where the temporal information may not align with the structure.

In each network there were 4 nodes $\{A, B, C, D\}$, and the contingency table describing the probability of each node occurring on any one trial can be found below in Table 3.3. Then each of the occurrences needed to occur at a particular time. It is in this sense that the conditions differed from one another. Condition 1 had no temporal information, so I will not be analysing it. The time associated with each event in each of the conditions is represented in Table 3.4.

Computers Activated	Probability
ABCD	0.512
ABC¬¬D	0.128
AB¬C¬¬D	0.128
AB¬C¬D	0.032
A¬B¬C¬D	0.200

Table 3.3: Generative distribution for data in Lagnado and Sloman [6].

Participants were then asked whether they believed each of the potential edges in the graph was present, they reported the proportions of responses yes to each of these edges as they varied according to experimental condition. A canonical version of these results can be observed in Table 3.5.

Strictly speaking, I handle only conditions 2–4 from Experiment 1 in Lagnado and Sloman [6], as the first condition has no temporal information and therefore cannot be analysed in this manner.

[ʊʊ] I interpret this to mean that they wanted the participants to infer the latent structure of the network as it existed then. An alternative interpretation is that the network was a complete network in which only some of the links were working at any particular time. I do not analyse the experiment under this interpretation.

[♣♣] In the course of studying this, at one point my model was performing quite poorly and making strong predictions that disagreed with the results reported in the paper. After contacting David

Condition	Activation Times			
	A	B	C	D
2	0	1	3	2
3	0	3	2	1
4	0	1	2	2

Table 3.4: Timing values for data in Lagnado and Sloman[6].

Edge	Condition 2	Condition 3	Condition 4
A→B	0.9600	0.5800	0.9200
A→C	0.1300	0.5400	0.3300
A→D	0.1700	0.8800	0.2900
B→C	0.7900	0.2100	0.7900
B→D	0.9600	0.3800	0.8800
C→B	0.3800	0.7900	0.5000
C→D	0.2100	0.3300	0.2900
D→B	0.4600	0.5000	0.4600
D→C	0.8300	0.7100	0.2100

Table 3.5: Canonical set of human responses for different edge proportions in Lagnado and Sloman[6]. This table's values do not agree with everything in the original publication, these are the correct values. The first node in a pair indicates the parent node and the second node in a pair indicates the child node.♣♣

3.12.3 A competing account: Local Prediction Learning[3]

My approach to modelling this problem requires using novel constructions beyond those described in the preceding continuous-time modelling cases, and beyond the methods standard for probabilistic modelling using generative models. These constructions differ in nuance from traditional approaches, and it would be easy to mistake my approach for being a variant on a standard approach. It will be easiest to comprehend the differences in my approach by contrasting it to more traditional generative approaches to modelling these data.

In their original paperLagnado and Sloman[6], Lagnado and Sloman did not propose a computational model that accounts for their phenomena. Their goal was to illustrate that participants use "both temporal and covariational information, with the former dominating the latter," or even that "temporal order can override sparse covariational information and lead to spurious causal inferences." For such a purpose, ANOVAs were sufficient. This is in sharp contrast with my account though – in which time plays a central role without which speaking covariation is nonsensical – they treat temporal information as something to be "traded-off" with covariational information. This is in line with the interpretation of the one model I know of that has been proposed to account for these experimental results, namely Wellen and Danks[3].

In Wellen and Danks[3], the authors suggest this experiment may be better modelled by an algorithmic-level, process model that guarantees finds a solution to the inference problem in a "computationally tractable manner". They propose a model that is a "hybrid between the associationist and rational paradigms in causal learning research." Theirs is an error-driven learning model called the "Local Prediction Learning" model, which takes its current expected predictions and updates its parameters according to the degree of error it finds between its predictions and the actual outcome. It assumes that edges are defined using a noisy-OR parameterisation, and that its parameters are analogous to that found in work on "causal power"[81,212,197,30]. It computes an expectation separately for each edge parameter and updates the parameter values accordingly.

It is worth noting here that despite the prominence of the role of time in the structure of the problem itself, the model will treat temporal information as an incidental feature of the

Lagnado directly I learned that my model was performing correctly and had been making accurate predictions; the numbers that had been included in the paper's results were incorrect. The correctness of the table in question was not crucial for the results of that paper and so it has gone uncorrected. However, I cannot think of a better point of evidence for the effectiveness of a modelling paradigm than being able to correctly predict that data reported in the publication you are modelling were reported in error to model.

underlying structure learning problem defined in terms of standard causal graphical models. That is, they saw it as a standard noisy-OR Bayesian network that had a known expected delay which is fixed at a delay of $1s$ (being a "natural delay unit").

This closely resembles the methods Lagnado and Sloman [6] used to construction their stimuli; the model structure(edge set) was fixed, the trial based (proportional) properties are those of a standard binary causal graphical model and were also fixed. The conditions vary along the temporal dimension, but the event times were only assigned *after* the occurrences were determined. In contrast, time will be an unavoidably central aspect to my model.

A WORKING HYPOTHESIS. To produce an expected prediction, rather than averaging over models as might be done in a Bayesian account, Wellen and Danks keep a working hypothesis of the graph on each update. This means that at any particular point and time there will be a set of "known" edges and a set of "potential" edges.

STRENGTH PARTICLES AND LEARNING. For each known and potential edge, there is a set of "particles" (in analogy to particle filters in approximate Bayesian inference [174]) that estimate the parameters associated with that edge (in the case of noisy-OR on a discrete trial with binary outcomes, this estimates of this parameter would be estimates of causal power [81,30]). The particles have a uniform indexed across all edges — i.e., if one edge has five particles, all edges(known and potential) have five particles and all particles with index 4 across all the edges will be used and updated with respect to the values of the other particles with that index. This is not quite the same as updating them jointly, because only a single edge (and a single particle for that edge) is evaluated and updated at any one step. However, because expectations about the occurrence of an event in a noisy-OR scenario, depend on the that update is based on the values of the other particles on the previous time-step, where dependence on the values of the other particles on the previous time-step ensures that there are no spurious effects from the order in which the updates occur. The learning occurs in terms of a general scheme by which one takes the difference between the expected and actual outcome, weight it by a learning parameter α and alter your particle's value accordingly. In mathematical terms (where $V^i_{C\to E}(t)$ is i-th particle's strength estimate for the edge from C to E at time-step j, α is the learning rate, $o_E(j)$ is the observation of E at time-step j, and $\mathbb{E}_{V^i(j-1)}[o_E(j)]$ is the expected value for E at time-step j according to the values of all the particles with index i at time-step $j-1$):

$$V^i_{C\to E}(j) = V^i_{C\to E}(j-1) - \alpha\Big(d_E(j) - \mathbb{E}_{V^i(j-1)}[d_E(j)]\Big) \tag{3.23}$$

Structure from strength. The task given to participants in Lagnado and Sloman [6] was to identify "whether they thought a connection was still working", providing a "yes" or a "no" answer. Thus, people However, the LPL model is defined relative to modifications of causal strength parameters. In order make a jump from strength estimates to structure judgements, Wellen and Danks take their particle distribution and run standard hypothesis tests (t-tests) against the null hypothesis that the edge has a strength parameter = 0. Given a p_{critical} parameter, if this test provides a p-value less than p_{critical} the edge is added to the list of known edges for the next iteration of the algorithm, and if a known edge's p-value is greater than p_{critical} it is removed as a known edge from the graph (it may later be reassigned known edge status, but on the next iteration it will be treated as only a potential edge). This strength estimation allows the model to make explicit structural claims despite being defined over modifications of strength values.

3.12.3.1 Time as influence

As stated before, whereas we see time as being crucially linked to the interpretation of data before covariational information can even be defined (if in most cases in terms of assuming a trial structure), Wellen and Danks [3] describe time as able to "influence contingency learning" via the "interpretation of covariational data". This is more in line with the role for temporal information presumed by Lagnado and Sloman [6]. In their words "The LPL model compares the observed temporal difference d_{E-C} between a potential cause and the effect to the expected temporal difference d_{typ}". Wellen and Danks [3] describe two models for doing this.

The first model is a deterministic model where "[i]f the learner expects the delay to always be d_{typ}, then the causal strength estimates update only when that delay occurs." They do not model this situation, but instead describe a case where "the learner expects the timeframe of the causal mechanism to be noisy, then the model *reduces* the saliency of C… as a potential cause of E…".

Specifically in their mathematical formulation, they treat the difference in delays as a way to weaken (or, more accurately, modify) the learning rate for a particular edge on the graph according to the difference between the expected delay (d_{typ}) and the observed delay (d_{E-C}) between those two events E and C where the edge under question is $C \to E$. They define the delay error as $d_{err} = d_{E-C} - d_{typ}$, and a particular (new) learning rate α' that "decreases exponentially as d_{err} increases: $\alpha' = \alpha \exp(-\frac{|d_{err}|}{s})$; s being a scaling parameter that deter-

mines how sharply α' drops off as $[|d_{err}|]$ increases".[♦♦] For their analyses, Wellen and Danks "set d_{typ} = 1 as the natural temporal delay between a computer sending a text message and receiving it would be one time-step."

A NOTE ON NOTATION. In describing the Wellen and Danks [3] model for incorporating time, I used their notation. As it is not compatible with the notation I have used elsewhere, I will not use their notation elsewhere. I merely chose to follow it to minimise any issues that could arise from the inevitable interpretation that occurs when translating between notational systems. Notably, while they use $d_{.}$ to refer to various kinds of "delay", I use d to refer to (various types and sets of) data and the notion of a delay (as a scalar quantity and not as a relative interval carrying a scalar measure, as denoted by τ) does not truly appear in my description.

3.12.3.2 Results from the Local Prediction Learning model[3]

DATA. Like any other model of the Lagnado and Sloman [6] experiment, the LPL model must address the problem that the stimuli were randomly generated for each of twenty-four participants. They chose to assume that every participant's dataset was exactly consistent with the expected value of 100 cases drawn from the generative model for the data, yielding: 51 cases with (A, B, C, D); 13 $(A, B, \not{C}, D)$; 13 $(A, B, C, \not{D})$; 3 $(A, B, \not{C}, \not{D})$; and 20 $(A, \not{B}, \not{C}, \not{D})$." In contrast, for my model, I will sample the 100 data points from the data generating distribution rather than use the expected value. For their model, they randomized the order of examples.

MODEL PARAMETERISATION AND EXECUTION. They ran the LPL model 1000 times, obtaining a single graph as the final working hypothesis for each model run (according to the t-test criteria described above). They assumed each edge had a generative noisy-OR relation and thus had a causal power ranging from 0 to 1. To estimate this they used 5 strength particles per edge, initially drawn from a truncated Gaussian with $\mu = 0$ and $\sigma^2 = .2$. They determined the learning rate α and their hypothesis testing threshold $p_{critical}$ by maximising fit (defined below) for condition 1, where there was no temporal information.[♥♥] This gave values of $\alpha = 0.1$

[♦♦] In their original paper, $|d_{err}|$ was written as d_{err}, which suggested a number of bizarre implications. It was confirmed through personal communication that $|d_{err}|$ was intended and in actuality what was used in their model.

[♥♥] Because of the lack of temporal information, I will not model this condition in my analysis below.

and $p_{critical} = 7 \times 10^{-5}$, which then were used to model conditions 2–4. They set $d_{typ} = 1$ and fit $s = 7$ to optimise their model fit across conditions 2–4.

Bayesian Model Comparison. They compare their model to a "standard Bayesian model of structure learning."♠♠ They modelled all possible graphs over the 4 variables giving each uniform prior probability, with a likelihood conditioned on the intervention on A as

$$\begin{aligned} P(d_j|H_i, \text{☞}_{t=0}(A)) &= P(b, c, d, t_B, t_C, t_D|H_i, \text{☞}_{t=0}(A)) \\ &= P(b, c, d|H_i, \text{☞}_{t=0}(A))P(t_B, t_C, t_D|b, c, d, H_i, \text{☞}_{t=0}(A)), \end{aligned} \tag{3.24}$$

taking $P(b, c, d|H_i, \text{☞}_{t=0}(A))$ from the parameterisation given to participants as each edge having a probability of .8 of working. Where $\text{☞}_{t=0}(A)$ indicates that there was an intervention setting node A to be active (implicitly ❖❖) In order to incorporate the temporal delay, they defined a distribution over for delay functions as

$$P(d_{E-C}) = \frac{1}{2s} e^{\frac{|-d_{err}|}{s}}, \tag{3.25}$$

but their actual likelihood is fairly complicated and based on taking into account the probability of different paths available on cyclic graphs✧✧.

♠♠ Given the number of challenges that have had to be overcome to develop the framework I discuss here, it is worth noting that a standard Bayesian model of structure learning with continuous time data of the sort implied by the exponential decay function does not exist. Wellen and Danks underplayed their originality in inventing this class of model within their larger project. Furthermore, it may have not matched their preferred model in terms of predicting people's exact judgements, being too certain in its conclusions, but it did quite well by the traditional standard of correlation r = .97 compared to the LPL model's r = .74. Still, I will hold myself to their standard.

❖❖ A more general notation for intervention is needed, as this will not be able to capture setting A in the case of a binary substitutive variable let alone a multinomial, real valued or vector value. However, accomplishing such a notation while also occupying too much space proves difficult.

✧✧ The likelihood model they developed is clever, and suggests ways of improving my own model, though those generalisations would need to apply even if I were *not* to consider cyclic graphs. At the time of my developing my model, I did not know of the Wellen and Danks [3] method for computing the likelihood, and thus did not apply it to my model. The gist of the changes would be to take the probabilities that I generated for each graph and treat that as a representation of observing the data underneath that graph being an accurate representation of the actual graphical structure on that trial. I interpret Lagnado and Sloman [6] literally then when they say there was a probability of an edge not working on a particular trial. That is, the final posterior for each of the graphs would be calculated by the expected

Performance Metrics. In evaluating their LPL and Bayesian models, Wellen and Danks use R^2, which is not equivalent to the correlation coefficient (r) squared (r^2) as is sometimes represented. Rather it is calculated across all the potential causal relations for which participants provided endorsements, as $R^2 = 1 - \frac{\text{SS}_{err}}{\text{SS}_{tot}}$, where SS_{err} is the sum of the squared errors between the model prediction and the participants' proportion of endorsement and SS_{err} is the sum of the squared errors between the mean-model (the model that predicts that each edge will be endorsed at a level equivalent to the average/mean value of participants' endorsements across all edges) and the participants' per-edge endorsement proportions. They interpret R^2 to be a measure of the proportion of variance that is explained by the model, saying that their model explains roughly half of the variance across the conditions ($R^2 = .47$).

Table 3.6: Table of R^2 values from Wellen and Danks[3].

	LPL Model	Bayesian Model
Condition 1	0.47	-0.03
Condition 2	0.40	0.81
Condition 3	0.46	-1.01
Condition 4	0.59	0.36
Overall	0.47	0.23

With this base-line model and set of predictions in hand, I can explain my approach to modelling this experiment.

3.12.4 Modelling hidden causes with continuous-time causal theories

The first step in building my model is to address the problem of the possibility that "effects" appear to precede their "causes". However, by carefully examining the structure of the problem it becomes clear that that is only apparent under a too simple model of the phenomenon. One straightforward way explain the data without allowing causal relations that reach back in time is to recognize that the observed variables are not the same as the causal variables. Or, put an-

value of the posterior for that graph under the probability that each one of its edges had a chance of not existing with probability 0.8. That is, the final judgement would be based off of a function of the calculated graph posteriors weighted by the probability that the graph would actually represent that. This would mean that graphs with more edges would have a greater number of contributions to their posterior, since this procedure only removes edges, it does not add them.

other way, the times at which variables spread their causal influence is not the same as the time at which variables are observed to occur. This is a more formally explicit way of capturing:

> "This means that it is possible for a message to be transmitted by a computer before it is displayed on its own screen(in the same way that you may pass on a virus before it becomes active on your own computer)."

Wellen and Danks [3] did *not* take this approach. They assumed that there is an expected $1s$ delay between causes and their effects, and that discrepancies between that expected delay and the observed delay reduce the amount learned from any particular trial. Thus, from their model's account of the phenomenon, when a graph claims $B \rightarrow D$, it is not saying that occurrences of A must precede B in order for A to cause B. If one sees a data-point that suggests effects precede their causes, one merely learns less from that data-point. Preceding a cause by $1s$ is formally equivalent to postceding it by $3s$ (if the assumption is that there is an expected delay of $1s$).

Merely penalising inaccurate delays misses some of the key aspects of the scenario described in Lagnado and Sloman [6]. People were not told that there was a fixed unknown time delay between various causes and their effects, and that errors in that delay could include effects preceding their causes. Rather, they were told that the computer could be infected before it was capable of displaying that fact. And it further misses the aim of the design of the experiment, that is to demonstrate a case where,

> More generally, the causal order in which events occur cannot simply be read off from the temporal order in which events occur (or appear to occur).

It is not that there is a noisy temporal relationship between cause and effect that allows effects to exist in the past of their causes, but that it is presumed that the events are causally efficacious at a time before their full consequences can be observed. Thus one needs to posit the existence of some entity or property that is capable of communicating causal effects (transmitting the virus) along the network, before the local manifestation of different effects (displaying the virus' message) on the screen.

There are at least two ways in which one could approach this problem of postulating hidden causes: modelling them as proper processes/nodes with events in their own right that need to be treated as part of the graphical inference problem, and modelling them as parameters of the variables that are observed that are associated with the amount of time that passes between the time when the variable is capable of communicating the virus and when the variable

displays its status on its screen. The former is more along the lines of hidden mechanism inference where the mechanism is cached out in terms of graphical structure, and work in this vein is in progress but has not produced results worthy of reporting here yet. Though I will, strictly speaking, be pursuing the latter in terms of the mathematics I use, this can be viewed as a means of approximating the full generative model produced by CTCTs.

I will model the incubation or latency intervals as parameters associated with the variable generated randomly on each trial. Doing this requires being explicit about three different kinds of events that were not distinguished previously. First there is the moment at which an entity is infected. Second, assuming a point cause, there is the moment at which the infected entity begins processes that will work over the hidden network to affect its children. Third is the display of the virus status on the screen. I will assume the first two kinds of events occur as part of the same event and therefore at the same time (i.e., becoming infected is the trigger for infecting others). Thus, we will need the latent interval to be defined relative to the moment of infection/infecting, which which for X is the unobserved $t_\star(X)$ and the display event will occur after δ_X that such that the observation time $t_\bullet(X) = t_\star(X) + \delta_X$.

It is worth noting that we can define all of these pieces entirely with respect to a single node X even though the original event itself may be generated by any of a number of parents of X. It might at first seem that δs in particular should be defined relative to the process that is caused, but I am treating them as being entirely internal to the node. The alternative would be needed if we wanted to provide a proper forward generative model, but doing so introduces a number of problematic complications[**]. By not requiring inference over a full forward probability model, but only a retrospective one, we can have a well formed modelling problem that is able to explain and induce observations and hidden observations without being able to predict observations (without access to the hidden observations).

[**] If this were a proper generative model, more details would need to be provided about how node X handles the case of being infected when it has already been infected but is still in the latent period before the effect takes place. For example, one possibility is that during the latent period defined by a $\delta_{X\to Y}$ until the infection of node Y is displayed after having having been caused by X at time $t_{X\to Y}$, if another cause Z were to produce an event at $t_{Z\to Y}$ on the Y before $t_{X\to Y} + \delta_{X\to Y}$, and the delay on that process makes it display sooner than the original display time (i.e., $\delta_{Z\to Y} < t_{X\to Y} + \delta_{X\to Y} - t_{Z\to Y}$) how one resolves this event depends on assumptions you make about the mechanisms in question. For example, it could be like fertilisation between an egg and a sperm cell, such that the features of the entity that is first to arrive are the only features that matter for the long term consequences of the process. Alternatively, it could be that it is more like a race, in which even if one entrant had a late start, if they are quick enough they can still cause the event of finishing the race first. One could also imagine combining their influences somehow in many different ways. Because I am not inferring the full forward generative model, I can avoid some of these issues.

3.12.5 A detective inference model for inferring hidden events

Under most standard generative models, you can generate hypothetical data(including hidden nodes) according to the distributions and parameters defined in your model. This will not be possible in the model I use as a stand-in for inference about the full generative model**. The key to why this will not work relates to the way that the incubation δs are defined and how non-events are handled. We need to ensure that all of the infection/transmission events ($t_\star$s) occur after $t_{☞(A)}$, but we do not know when any of the infections actually occurred. We only know the times of observations ($t_◉$s), for those events that did occur. We know that the $t_\star$s will be the time at the beginning of the interval that ends in the observations($t_\star = t_◉ - \delta$). But, given the constraints on when $t_\star$s can occur, the support space for δs will depend on when the event is observed. An event that is observed that later in time has a wider range of values that its associated δ can inhabit. This means that there is no fixed value that determines the distribution of δs irrespective of the observed data points; such a fixed value is required to accommodate this kind of filtering in a standard generative model.

And here is where we can use insights from human inferential abilities to look at the problem in a slightly different light. If we look to the medical examples that I have been considering so far, often they rely on assuming that some unobserved but real event occurred at some point before the event to bring about this event. John Snow did this when he searched out a well as a point-cause of cholera being spread throughout Soho, given his theory that it was transmitted via the water supply[213,214]. Furthermore, for examples that did not fit his expected pattern (such as an old woman who lived far away who died of cholera) he attempted to identify how their sickness can be explained through his theory based on learning additional information about hidden events (the woman liked the taste of water from that particular well so she paid someone to bring her water from it despite its being far away). From the observations he reasoned backwards to detect the event that lead to the observations. Relying on his theory to supply candidate mechanisms he revealed the hidden structure of the causal system with this kind of reasoning. It is unlikely that he would claim that in doing so he would uncover all of the events that could have potentially been generated by the mechanism, meaning that his inference would fail to fully characterise the generative process of the system. However, he could treat his detective work as a kind of approximate inference about events caused by the underlying generative model. We need a probability model that can perform exactly this type

** Though I have a functioning framework for deriving such a forward sampling model, the programming problem has proven as challenging as deriving the formal framework. A version of this that does not deal with base-rate causes is introduced in Appendix C.

of inferential short-cut.

Detective inference is the approximation of a generative probability model that may include assumptions that violate some aspects of the generative model and may not be generative probability models themselves. Given some data, the inferential process' job is to detect what unobserved events occurred to bring about the observed events, and thereby infer after the fact what the structure of the system was that produced the particular observations. This is in contrast to the generative model that would also propose the existence of hidden events that did not produce any particular observations within the observation period. The detective will still need to state how likely it is that particular entities did not occur, but it will not reason forward from the intervention and postulate all of the potentially hidden events that would occur as a result of the intervention, but it will reason backward from the observations that occurred to identify their source.

3.12.5.1 Imagining the causes of unexpected events and calculating their likelihoods

> *"See the value of imagination," said Holmes. "…We imagined what might have happened, acted upon the supposition, and find ourselves justified. Let us proceed."*
>
> Doyle [215], in *The Adventure of the Silver Blaze*

A detective is not an oracle, they can look into the past, but not necessarily the future. Accordingly, a detective inference model may not be well defined to provide forward generative predictions about what events would occur were the stochastic system to be sampled from again. There may, in fact, be a generative model closely analogous to the detective model that is capable of obtaining forward samples, but a detective inference model does not need to adhere to the constraint that its "prior" distributions needs to be fully defined a priori. Specifically, my inference model can constrain δ distributions on the basis of observations that tell us when and whether any event occurs. These are not true prior distributions because we cannot postulate hidden events that occur before the intervention $t_{☞}$ that initiates the events and observation period.

A detective does not engage in exhaustive search, they generate hypotheses for the observed events and investigate the consequences of those hypotheses. Thus, detective inference models will be most applicable to point process models of the kind I report here and event sequence models, where there is an unavoidable asymmetry between events occurring (points) and not

occurring (periods). Or, to put it more precisely, a detective inference model may be most applicable when events occur that are a violation of what would otherwise be expected in the context of the rest of the causal system. For example, Sherlock Holmes cites "the curious incident of the dog in the night time" as evidence, where the curious incident was the dog's silence when barking would have been expected had a stranger been present at the crime scene [🐾]. In my case, my model will only identify the source of events as they are comparatively rarer than the periods of time during which no events occur.

With samples from the δ distributions and the observation times, my detective inference model can generate postulated samples of when hidden events $t_\star$s occurred. And, with $t_\star$s in hand, I can use the causal theory to to evaluate the likelihood of the total simulated event set. Because the theory builds the prior for my models including constraints on the values δs and $t_\star$s can take given $t_◉$, this can be seen as a variety of empirical Bayesian inference in the sense that my prior is dependent upon the observed data rather than being fixed. However, unlike standard empirical Bayesian inference[216], there is no uncertainty about the parameter being "inferred" — it is a hard constraint that the δ cannot be greater than its corresponding $t_◉$ as that would make $t_\star$ negative, meaning it occurred before $t_☞$, which is explicitly disallowed. Additionally, unlike standard empirical Bayes which aims to estimate a parameter governing a series of repeated observations of variables with static distributions, I will be sampling from this distribution in order to estimate simulated unobserved events in a dynamical system. And I do this because the exact timing of those hidden events determines which parts of the dependency structure are relevant to computing the likelihood underneath any one model.

Given a set of postulated $t_\star$s, we will be able to define their relative likelihoods of occurrence as well as the likelihoods of no event having occurred during the times between occurrences. This includes those periods of time during which we presuppose no hidden event occurs because no observed event was seen to occur. These likelihoods will be defined relative to the models generated by CTCTS, and – in addition to the simulated events – will depend on parameters that can be obtained from standard sampling procedures, much as they would be in a traditional generative model.

[🐾] The dog's particular silence was only notable because there was another event that would have otherwise been expected to trigger a barking event. This triggering event allows the silence to be a specific unexpected event in contrast to most of the time when the dog is also not barking, but which did not constitute evidence for the case. This harkens back to the Mackie[149] K machine in which a failure to produce a chocolate bar is tied to a particular instigating cause and concurrent particular failure.

How does a detective inference model differ from the generative model? Forward inference models will be able to make stronger and more accurate claims about hidden events than the retrospective detective models. For example, forward models will not make the assumption that if no observed event has occurred that no hidden event occurred. Additionally, one notable feature of real-world detectives is their postulating a specific hypothesis regarding parts of the causal mechanism rather than marginalising over uncertainty. My inference procedure does the same. It may be the case such that when detective inference models are needed, this retrospective single hypothesis generation process will be the most tractable or interpretable. A Monte Carlo estimation approach can allow multiple samples over which estimation could occur, but to say how exactly that would work requires a much more extensive formal analysis than I will be engaging with. A general solution may be impossible, given the difficulties of specifying anlaytic forms for the joint likelihood of hidden events in entangled causal systems like those I am analysing. For my retrospective modelling, it suffices to have a single sample or hidden event hypothesis for each observed event. A forward model would also be able to predict the occurrence of hidden events that have no observable counterpart on the grounds that such a counterpart would not be seen within the boundary of observation (e.g., because observation was halted after 4 seconds). A retrospective model like this could not postulate those hidden events because they would be unable to identify a specific $t_{\bullet}$ from which to calculate a $t_{\star}$ given the appropriate δ since $t_{\bullet}$ by definition was unobserved.

This aspect of this section's model is perhaps its most unique feature, both in the context of the problems I am discussing and in the larger probabilistic modelling literature[△△]. We expect detective inference models will be of use in situations analogous to those where real-world detectives would traditionally be needed. Retrospective models will be of most use when events have occurred due to unknown causal mechanisms with observations believed to have

[△△] In fact, Jessica Hamrick has pointed out that this detective model – were it to be treated as a full model of the situation and not merely an inferential approximation to the generative model – can be interpreted as flipping the reversing the direction of the arrows to be contrary to the causal direction of effects and then this distribution should be well defined. This arises because then the observation is taken as given, and from that you can jointly reason about the observations' "effect" on its cause in terms of the observation determining the value relative to which the other variables will then compute likelihoods. However, to properly describe the prior distribution on the times may require an arrow from the intervention directly to the observed effects as well, to be able to compute the time relative to the point intervention. Otherwise it is unclear how one is able to define the range of potential values the $t_{\star}$ and thus the δ could take on, which would not solve the problem of having undefined probability distributions for generating the δ. It is not clear if the assumption of a universal time metric and setting the time metric to be 0 at the point of intervention solves this problem or merely hides the role of the structural syntax inside the semantics of the time metric.

a specific causal origin. They will arise most often when there are many potential details and ancillary considerations that would beguile a completely forward model in the complexity of considering all the predictions afforded by a closely related theory.

3.12.6 Defining the graph set: supergraphs and base-rate smoothing

There are 4096 (2^{12}) possible directed graphs (including those with cycles, but no self-loops) that can be generated from the 4 nodes $\{A, B, C, D\}$. However, because the participant intervenes on A and any node can be infected by the virus only once, all edges that lead into A can be removed from the set. This leaves us with 9 potential edges ($\{(A \rightarrow B), (A \rightarrow C), (A \rightarrow D), (B \rightarrow C), (B \rightarrow D), (C \rightarrow B), (C \rightarrow D), (D \rightarrow B), (D \rightarrow C), D\}$) and thus 512($2^9$) possible graphs.

Additionally, by intervening on A it is possible to affect B, C, or D, which suggests that there is a directed path from A that reaches each of B, C, and D. I use this information to reduce this set of 512 to a set of 304 graphs that meet these criteria. However, we need not do so if we allow for a non-zero base-rate of occurrence[¤¤]. Were this to have made a large difference, I would present analyses for both the complete graph set and the reduced graph set (but they lead to roughly the same results). Rather than enumerating these here, it is easier for the interested reader to go to the cbnx(Causal Bayesian NetworkX) repository on GitHub where there is code that will allow these graphs to be generated programmatically by adhering to these conditions.[¢¢]

3.12.6.1 Defining a supergraph given a graph set.

If we are going to compare the different graphs using a Bayesian inferential procedure, we will need to define a prior over those graphs. It is common when doing this to allow for sparsity

[¤¤] Note, allowing non-zero base-rates violates the instructions Lagnado and Sloman [6] gave to participants stating there were no hidden causes. One would want to exclude some of those graphs, as they would be infinitely unlikely to produce the data if they were not allowed to have hidden causes/base-rates of activity.

Because my model presumes the existence of base-rates of activity (see minor section 3.12.6.2), there is no reason to artificially limit ourselves to those graphs. Indeed, if I can predict human judgements while computing over the larger graph I have shown that my model is unable to be swayed by the existence of the other graphs to make unrepresentative predictions.

[¢¢] If one wishes to characterise what I describe later as latency parameters as true hidden nodes, this picture becomes both more and less complicated. For further details I direct the interested reader to the Appendix B, Appendix C and the hidden structure inference GitHub repository where there is code for implementing this more complicated version of the model without including any base-rates.

considerations to play a role, that is, where you will penalise graphs with more edges compared to fewer edges. However, to describe there being more or fewer edges, it can often be useful to describe this in terms of a supergraph, which – in analogy to superposition – consists of a union of all the edges(and therefore nodes) of the graph-set in question [$$].

This process allows building a system for describing a set of graphs first, and then using the supergraph as a part of the definition of a prior for the graphs. This avoids the problem of needing to define the prior's value for potential graph structures in the same terms generative model that defines the possible graph structures. Aside from convenience from a programming perspective, it may allow mutually incompatible graphs (graphs that do not share the same sets of nodes) to be brought in comparison with one another as part of the same theory[€€]. The intent is to define a prior across all the graphs by providing a unnormalised probability distribution that is a function that supports each of the graphs as input, and then to normalise after the fact.

In my case, for modelling Lagnado and Sloman [6], we will be dealing with homogeneous edges and nodes (all edges are of the same type and all nodes are of the same type), though that is not strictly necessary. For a set of graphs $\mathcal{G}$, let $\hat{\mathcal{G}}$ be the supergraph, which is formally defined as the graph containing union of all edge sets (and therefore all nodes) in all of the graphs that are parts of $\mathcal{G}$:

$$E_{\hat{\mathcal{G}}} \equiv \bigcup \{e = (X, Y) \in E_G | \forall G \in \mathcal{G}\}.$$

I will use this to consider the absolute number of edges in a graph in comparison to the supergraph,

[$$] It is not clear how easily this notion will generalise to nonparametric situations with infinite numbers of potential entities. However, given that models of this sort have been proposed in the theory based causal induction framework [1] that treated hidden causes as being invoked from an infinite set as needed according to a version of the Chinese restaurant process [217,218], I believe that it is compatible with such a notion. In fact, I believe it may even be compatible with non-exchangeable nonparametric probability distributions (probability distributions of infinite sets of variables that are not guaranteed to be independent of the order of steps taken when building the graph as described in a generative model) such as the distance dependent Chinese restaurant process [219]. That said, I believe it would require different ways of accounting for finite and infinite sets of potential hidden relations that take these subtleties into account. Here I only deal with finite numbers of homogeneous sets of potential edges between finite sets of homogeneous nodes.

[€€] Presumably this method would only be pursued if these different graphs are intended to model at least some data in common between them, as otherwise it is difficult to comprehend the need for a prior that spans their structures.

$$\#(\hat{\mathcal{G}}) \equiv |E_{\hat{\mathcal{G}}}|.$$

This is not the only way you could use a supergraph. For example if you had different edge types, you could count those individually, or you may want track features that indicate which parts of a causal theory need to be invoked in order to generate the different nodes or edges in the supergraph.

3.12.6.2 Base-rates and smoothing over graphs

I will need to introduce a base-rate of events occurring for computational and inferential reasons. The reasoning behind this modelling choice is most clear when illustrated by example.

Because I will be using a detective model with many different graphs, there is the possibility that on any one retrospective sample, the way that events happen to line up is valid according to one graph, but invalid according to another. This is not a constraint on the $t_\star$s in relation to the initial triggering intervention (i.e., it is not because the $t_\star$ is negative according to $t_eye-\delta$), but the order in which cause events occurred as it relates to the graph structure. For example, if the first graph (G_1) under consideration was $A \to B \to C \to D$, but according to a detective simulation $t_\star(D) < t_\star(B) < t_\star(C)$ the likelihood of observing those values under the no base-rate of events assumption is 0. Even if you modified the graph (G_2) to add the edge $D \to B$, it would still be 0 as there would be no way for D to be triggered initially because C is its only parent and it still occurs after D. In fact, if we adhere to no base-rates, only graphs containing the edge $A \to D$ (G_3), would have non-zero likelihoods.

This has a few consequences. First, in the context where there are many trials being evaluated (in my case, 100 trials) the probability that such an order will happen for at least one of those trials on each sample for at least some graphs is high. That means that the kind of sampling I have described using in our detective inference model will be inefficient and that even those samples that happen to work for some graphs on some trials will be entirely cancelled out. Second, it means ignoring relative graph performance in accounting for partial datasets. That is, a graph that cannot explain *any* of the simulated data are treated as equivalent to graphs that explain part of the data well but some parts poorly. To make this concrete, the intuition is that G_2 is a slightly better graph than G_1 for accounting for the above example of explaining why $t_\star(D) < t_\star(B) < t_\star(C)$, because it at least accounts for the order of events if D were able to spontaneously occur. But if we assume there is no way for D to spontaneously occur, G_1 and G_2 are treated equivalently (with both of them giving 0 likelihood

for the data). Third, there will be sets small graphs that require mutually incompatible times, such that if one of them is to get a non-zero likelihood for a particular data point, there will be other graphs with the same number of edges that will only be able to get a zero likelihood. For example, the graphs $A \to B \to C \to D$ is incompatible with $A \to B \to D \to C$, since the former requires $t_\star(C)$ to precede $t_\star(D)$, whereas the latter requires exactly the opposite. Especially when one takes into account that there is no a priori reason for either of them to be impossible, meaning that a priori there is a possibility that either simulated data patterns occur, meaning for any one sample, a different set of graphs will give the data o likelihood. With many trials taken together we multiply their likelihoods making one o likelihood value propagate to be the likelihood for that graph across all the data points. This means that the more data points, there are more small graphs that are expected to be eliminated from consideration. Finally, because of the previous reasons, this method results in biased estimates that favour graphs with many edges, because they are capable of accounting for almost any order of hidden events. Worse, it is not that it gives dense graphs have a high likelihoods, but only that they are the only graphs that have *non-zero* likelihoods. And it's worth noting that dense graphs would intuitively not be expected to have large likelihoods, because they do a poor job of explaining why so few events occur(since all links are generative a denser graph would expect more activity). But, there is no similar absolute penalty for doing a poor job at predicting that events will fail to occur.

But what would having a nonzero base-rate actually amount to?

Too large of a base-rate will swamp the influence of graph structure on the likelihood. The stronger a base-rate the more equally likely any order of event times becomes. As the base-rate becomes exceptionally large relative to the model's structural rates, the biggest differentiator between different data values will be the absolute amount of time that passed before observation. Another way to view this is from the Poisson perspective, since I will be modelling data as the first arrivals from various point processes. As discussed before the likelihood that an event occurred due to any particular process of a set of processes that have been superposed upon one another is defined as the proportion of the net intensity provided by the process at the time of occurrence. A large base-rate will be constant at all times and is definitionally contributing more to the net rate of occurrence than the structurally induced processes. As a result, the probability that any event arrives due to the base-rate process is going to be quite large, which is a strict violation of the experimental conditions.

With a small base-rate, events that could not otherwise occur(according to graph structure)

can still be simulated without incurring an infinite penalty[⊠] to the log-likelihood of the data under that graph structure. Instead, it suffers a large, but finite penalty for having events that are only comprehensible in terms of the base-rate.

In this sense, we can see including a base-rate as a manner of smoothing out the predictions of our graphs, and doing so in a principled manner. There may be other ways to perform that smoothing (for example giving each graph a nominal amount of unnormalised probability and renormalising analogous to Knesser-Ney smoothing in natural language processing[220]), but this has a specific semantics that not only can be interpreted in terms of our causal framework, but that interpretation is straightforward, common and reasonable. That a base-rate exists due to potential unknown causes is an inevitable uncertainty faced in real life, and the functional form of genuinely unknown causes has no reason to vary over time, as that variance would suggest that something about the cause was known, making a flat base-rate appropriate. Including a base-rate is a standard (and often necessary) assumption in defining probabilistic causal models(see, for example Griffiths and Tenenbaum [1]).

This principled manner of smoothing our likelihood under different graphs by using a reasonable causal model that includes a small constant base-rate allows us to directly address the above problems that arise from having zero base-rates. The first strength is that with any nonzero base-rate all event sequences are possible even if some are only able to be explained via the base-rate according to any particular graph. This allows graphs that make good partial predictions to be given a higher likelihood than than graphs that fail to predict any events altogether. Smaller graphs that would be entirely unable to contribute final judgements because they would almost always have o posterior probability can now have nonzero likelihoods. Mutually incompatible small graphs will be able to have nonzero likelihood for the same data points, allowing us to get better estimates of the total set of the graphs' likelihoods in comparison to each other based on the Monte Carlo inference. And finally, dense graphs and sparse graphs will both have non-zero likelihood, meaning their comparison will not be intrinsically biased because of the Monte Carlo sampling mechanism for estimating the hidden event times.

[⊠] Automatically setting a likelihood to o according to one data point is equivalent to an log-likelihood $\ell(x|G) = -\infty$.

3.12.7 Parameterising causal theories

3.12.7.1 Perfectly probable events: interventions and one-shot events.

One of the important features needed to model this is the ability to represent quantities that because of the structure of the problem are perfectly probable in the sense that accounting for these facts cannot reduce the probability of the hypothesis that includes those facts. The two perfectly kinds of probable events we need to consider are point interventions and one events that can only occur once, the former for the occurrence of an event the latter for the non-occurrence of events over a period time. In continuous time, we model a point intervention by positing a point event that is taken as given (i.e., we do not need to compute its likelihood). Furthermore, because events are going to be defined relative to the point of intervention on series of trials, doing so will also be defining the time-scale for the rest of the processes relative to that intervention. One-shot events will be modelled as before, where once an event occurs it cannot occur again, which can be thought of as a perfect cancellation of all activity after the first event has occurred. In the context of Lagnado and Sloman [6] this means that once a computer is infected, it cannot become infected again. By extension it means that if a process is going to have a causal influence on other processes, the event that matters for causing other effects will have to be the first event to occur (as no other events will occur to transmit influence).

3.12.7.2 Meta-parameters

In order to define the sampling regimen for my model parameters, I need to define distributions for those parameters. Because I am not treating them as proper hyperparameters, but they are parameters that go beyond the parameters we will be sampling for our processes, we will refer to these as meta-parameters. Some of these meta-parameters – such as the expected incubation time – I can set using the information provided in the stimulus description. For others, I need to search for well-fit values, but will only consider values in scales that are reasonable given the background knowledge provided with the scenario.

3.12.7.3 Graph Metaparameter: Sparsity prior

I will use a sparsity prior over our graphset, where a graph is computed based on the number of edges it has compared to the number of edges in the supergraph $\hat{\mathcal{G}}$. If I take the sparsity

parameter to be θ_{sparsity} or θ_s, this gives us the unnormalised prior distribution

$$p(G; \hat{\mathcal{G}}) \sim \theta_s^{\#(E)}(1-\theta_s)^{(\#(E;\hat{\mathcal{G}})-\#(E))}, \tag{3.26}$$

for each graph.

I need to then normalise it to be a proper probability distribution by summing the values over all the graphs in $\mathcal{G}$ and dividing by that sum. One of the advantages of working with an unnormalised version of this distribution with a probability distribution defined relative to the supergraph $\hat{\mathcal{G}}$, is that it allows us to use the same effective formula for building a prior that adapts to the graph-set in question. For example, if I include a base-rate of activity to explain the occurrence of otherwise impossible events, I can consider both potential graph-sets (the well-fit graph set and the full graph set) and because both graph sets are defined relative to the same supergraph $\hat{\mathcal{G}}$ we can do so using the same parametric family of priors. Indeed, prior to normalisation, graphs have the same value under both circumstances.

3.12.7.4 Process Metaparameter: Latency/Incubation.

For every observed event, I will postulate that there was some time interval of length δ between the time of infection $t_\star$ and time of observation $t_\text{◉}$. This will have a fairly unique distribution because we never directly observe $t_\star$ but have strong conditions on when it can occur and we do observe $t_\text{◉}$. Because we know that the events in question must have occurred after the intervened on event, that means that there is an upper bound on the length of δ but that that upper bound depends on the actual occurrence of the event. If we take the time of intervention to be the starting point from which other times will be measured $t_{\text{☞}A} = 0$, then our $delta$ cannot make the cause time occur at a negative time. But we want to be able to sample a delta for every event, and, because we will define the data arriving on each trial as independent, we can resample δs for a particular trial according to the same distribution [⁘] for every observed

[⁘] It's worth noting the difference between a truncated distribution and a distribution that is not truncated but is treated as a standard distribution for whom some of the values are effectively filtered according to some logical conditions that give 0 likelihood to certain values of the parameter. The truncated distribution is treated as if it did not have support over that space. Therefore, were we to use Monte Carlo rejection sampling (which we are effectively doing but with a single valid sample in my case) from whatever distribution that we are truncating and resampling from, we would need not rescale the estimate based on the number of times that the sample is rejected. We do not need to account for the number of times that the sampling failed, that is merely an implementation detail of how to sample from the distribution. If we wanted to treat this distribution as a *filtered* distribution made to have o likelihood for some of its values – rather than a truncated distribution – we would sample in

event as a truncated exponential random variable with mean θ_{latency} or θ_l and truncation point equal to $t_{\text{◉}}$. Because I am including an observation in the definition of the prior, this is akin to a empirical Bayesian approach[216], where we inform our prior based partially on the data itself (since otherwise we are unable to specify the truncation parameter). So given a θ_l and $t_{\text{◉}}$ we can sample

$$\delta \sim \text{Exp}(\theta_l), 0 \leq \delta \leq t_{\text{◉}}. \tag{3.27}$$

$$p(\delta; \theta_l, t_{\text{◉}}) \propto \begin{cases} \frac{1}{\theta_l} e^{-\theta_l \delta} & 0 \leq \delta \leq t_{\text{◉}} \\ 0 & \text{otherwise.} \end{cases}$$

On the grounds that the hidden mechanism should be able to occur entirely within one time step (given the occurrence of events at 1 time-step after the intervention), that means there needs to be enough time both for the original process to produce its effect and for the latency to take action. Thus, I set $\theta_l = .5$ allowing, in expectation, 2 events to occur before the first event.

Technically, because $t_{\text{◉}}$ is defined for each observed node and can vary, the distribution for the different δs for different nodes will depend on how the actual occurrence of that node on a particular trial. I do not need to compute a δ for any nodes which did not occur, though if you interpret $t_{\text{◉}}$ as being $-\infty$ you can compute it and subtract it without concern.

Process Metaparameter: Base-rates. As discussed above (see minor section 3.12.6.2), I will presume that each process has a base-rate of occurrence of generating events. I will notate the rate at which events occur on node X as $\lambda_{\varnothing(X)}$ (or as $\lambda_{\varnothing}$ when it is clear that the rate under consideration is only concerning one node, making the X redundant). I will assume that the base-rate is constant at all times, meaning that it does not need to be indexed by time. Each base-rate is sampled using a base-rate metaparameter $\theta_{\text{base-rate}}$ or θ_b as

exactly the same way to get at least one valid sample, but would count the number of times failure occurred and penalise the likelihood as having been created in expectation for that trial's data. By treating it as a expected value for the likelihood, where each of the k failed samples count as having 0 likelihood for that Monte Carlo estimate, then you effectively are dividing the likelihood by $k + 1$ to account for the failures. This is closer to a standard Bayesian approach(in contrast to the former empirical Bayesian approach) that just allows more efficiently sampling a posterior when some region of the parameter space is guaranteed to have 0 likelihood. This version of the model may be able to be described in terms of a pure forward generative model.

$$p\left(\lambda_{\varnothing(\cdot)}\right) = \frac{e^{-\frac{\lambda_{\varnothing(\cdot)}}{\theta_b}}}{\theta_b}, \lambda_{\varnothing(\cdot)} \geq 0.$$

3.12.7.5 Causal Relations and Metaparameters

All the links that I am considering will be generative so I will only need to parametrise the induced processes as superposed FPPs (and I do not need to consider activations that induce thinning processes). This requires defining a rate function relative to the onset of the FPP. Thus, it will need support over $\mathcal{R}^+$, because it will only be defined in relative terms. Additionally, because these processes will be defined relationally, there will need to be one for every edge that exists on a graph and every such process will be indexed according to the two nodes on that directed edge. So, for example, if X occurred at time s $X \rightarrow Y$ would induce an FPP governed by rate function $\lambda_{X \rightarrow Y}(\tau = (t - s)) = f(\tau)$, where τ indicates how much time has passed since s given the absolute times s and t.

I will consider two kinds of intensity functions: one with a constant rate that does not decay over time and an influence that does decay over time according to an exponential decay function as in section 3.10 and section 3.11.

An intensity function with a constant rate would be

$$\lambda_{X \rightarrow Y}(\tau) = \psi_{X \rightarrow Y}, \tau \geq 0.$$

An intensity function with an exponential decay [⁾⁾] can be described in terms of how strong of an effect it produces and how quickly that effect decays parameterised by a maximum intensity and a decay rate:

$$\lambda_{X \rightarrow Y}(\tau) = \psi_{X \rightarrow Y} \exp(-\phi(\tau)), \tau \geq 0.$$

You can see the constant rate as the limit of the exponential decay function as the limit where the decay rate goes to 0(at which point the process would not be appropriately characterised as a finitary Poisson process).

INTENSITY. The intensity parameter $\psi_{X \rightarrow Y}$ will be a max-intensity parameter signifying the maximum instantaneous influence exerted by the cause X on the effect, it will be defined for

[⁾⁾] If you were to use other functional forms for the decay (e.g., a more general gamma functional form) a maximum intensity may not be an appropriate stand-in for the notion of the strength of a cause.

each edge. It will be distributed according to the $\theta_{\text{intensity}}$ or θ_i metaparameter as

$$p(\psi_{\cdot\to\cdot}) = \frac{e^{-\frac{\psi_{\cdot\to\cdot}}{\theta_i}}}{\theta_i}, \psi_{\cdot\to\cdot} \geq 0.$$

I set the mean of this exponential random variable to be $\theta_i = 1$.

Decay. The decay parameter ϕ can take on positive real values $R^+ \equiv [0, \infty)$, however if it takes on the value of 0 the likelihood will be of a slightly different form than if it is a strictly positive real number $R^{+\setminus 0} \equiv (0, \infty)$. A decay parameter of o is equivalent to a statement that the influence of the cause never weakens due to the passage of time (though this does not prevent other influences from affecting the effect). Assuming a non-zero decay parameter is wanted, then it will be defined relative to a θ_{decay} or θ_d,

$$p(\phi_{\cdot\to\cdot}) = \frac{e^{-\frac{\phi_{\cdot\to\cdot}}{\theta_d}}}{\theta_d}, \phi_{\cdot\to\cdot} \geq 0.$$

If the induced rate is constant, the decay is set to 0 and we do not need a distribution to represent it (or it is represented by the Dirac δ).

In what follows, I will report results for both decaying and constant models.

On modelling the cancellation of single-trial events. The model that I am describing has no other way to encode event cancellation in order to have well formed likelihoods. Simply including a multiple that acts as a "filter" (as I describe prevention in subsection 3.7.3) scales the max-intensity parameter, effectively changing the prior without affecting the structure in a meaningful way. Other methods should be pursued that can more accurately account for cancellation as a discrete occurrence on a single trial. [MM]

[MM] In fact, I have since learned of a method that allows us to encode event cancellation by applying cancellation distributions as part of our prior that equates graphs with those graphs that could be in a supergraph relationship with the graph. However, there are major implementation and interpretation details that make this not as easily applicable to the kind of model I posit here. In particular, if we were to cancel only one of those edges (i.e., compute the probability of the graph partially in terms of another graph with one edge removed) we may still have a graph that has its probability computed partially in terms of the cancellation of some of its edges. This recursion seems strange, and to instead necessitate a model that gives a different meaning to the notion of active and non-active than edge existent and edge non-existent. In fact, my framework may be exactly perfect for reasoning about such a system. The current model computes the likelihood of the data under the graph as though it were the true underlying relational system. If we need to think of the network as itself existing as a distribution over several true

3.12.8 Likelihood model

In the final analysis, I will need to aggregate data over each trial, but because trials were to be treated as independent, I can simply multiply these likelihoods (add their loglikelihoods) individually. There were K total trials so this means:

$$\mathcal{L}(\mathcal{D}|G_\alpha) = \prod_{i=0}^{K} \mathcal{L}(D_i|G_\alpha)$$

And then I can divide up the likelihood for a single trial into a number of subparts. Remember, for data from a single trial D_i, I have to consider two cases: those nodes that were never observed and those nodes that were observed.

Those that were never observed, I assume with the detective model, have no $t_\star$. These nodes still need to be taken into account as this non-occurrence of the $t_\star$ is indicative of the absence of an event that could have resulted from parent events but did not. This contribution is $\ell_0()$:

$$\ell_0(D_i) = \sum_{\substack{X \in B,C,D \\ \{X_i=0\} \subset D_i}} \left(-\lambda_\varnothing(t_{\text{fin}}) + \sum_{\substack{p \in par(X) \\ \{p_i=1\} \subset D_i}} \left[-\frac{\lambda_{p \to X}}{\phi_{p \to X}} \left(1 - e^{-\phi_{p \to X}(t_f in - t_\star(p))} \right) \right] \right), \tag{3.28}$$

Then, I need to take into account the nodes that were observed, which contribute two parts to the likelihood. I are going to be calculating the likelihood from the hidden time of infection ($t_\star$) not the time of observation ($t_\text{◉}$) — our detective method for computing the time of infection already took the probabilities linking those by sampling. The first part of the occurrence at the time of the occurrence $\ell_{1^+}()$ and the likelihood of the non-occurrence up until the point of occurrence ℓ_{1^-}. These likelihood components only consider those parents whose $t_\star$ preceded

underlying relational systems (defined by the edge cancellation probabilities and subgraph relations), then the posterior of the network is just the aggregated contributions of its subgraphs weighted by the edge cancellation probabilities.

the event's $t_\star$, and are

$$\ell_{1^+}(D_i) = \sum_{\substack{X\in B,C,D \\ \{X_i=1\} subset D_i}} \log\left(\lambda_\varnothing(X) + \sum_{\substack{p\in par(X) \\ p_{trial}=1 \\ t_\star(p)\leq t_\star(X)}} \left[-\lambda_{p\to X} e^{-\phi_{p\to X}(t_\star(X)-t_\star(p))}\right]\right), \tag{3.29}$$

$$\ell_{1^-}(D_i) = \sum_{\substack{X\in B,C,D \\ \{X_i=1\} subset D_i}} \left(\lambda_\varnothing(X) t_\star(X) + \sum_{\substack{p\in par(X) \\ p=1 \\ t_\star(p)\leq t_\star(X)}} [-\frac{\lambda_{p\to X}}{\phi_{p\to X}}(1 - e^{-\phi_{p\to X}(t_\star(X)-t_\star(p))})]\right). \tag{3.30}$$

I do not need to take into account the time after $t_\star$, because the probability of nothing happening on that node after that is 1.

3.12.9 Results and Analysis

The results from the CTCT model with decays are both quantitatively and qualitatively promising, as can be seen in Figure 3.13. Qualitatively, we have a good fit with the data with an overall correlation between the model predictions and human judgements of $\rho = 0.8775$. Quantitatively, I find that roughly 70% of the variance is explained ($R^2 = 0.7031$).

However, I do find the degree of success varies between conditions (as can be seen in Table 3.7). In particular, this seems to derive from the relatively large values given by participants to links between the nodes that they intervened on (A) and other nodes. Participants did so despite the time series being ($A = 0, D = 1, C = 2, B = 3$) for events when the did occur. This suggests that CTCT model discounts the temporal distances strongly, it seems that people did not do so to nearly so great a degree. Further evidence for this can be seen if I look at the distribution of R^2 values for different sparsity parameters. Unlike the overall, condition 2 and condition 4 curves, the condition 3 curve finds its R^2 value growing monotonically because the errors it makes are consistently underestimating the number of edges that present.

I find similar results for the CTCT model with no decay. Again the model performs well on both quantitatively and qualitatively metrics overall, as can be seen in Figure 3.15. The overall qualitative and quantitative fits are comparable to the delay model: $\rho = 0.$ and $R^2 = 0.6972$.

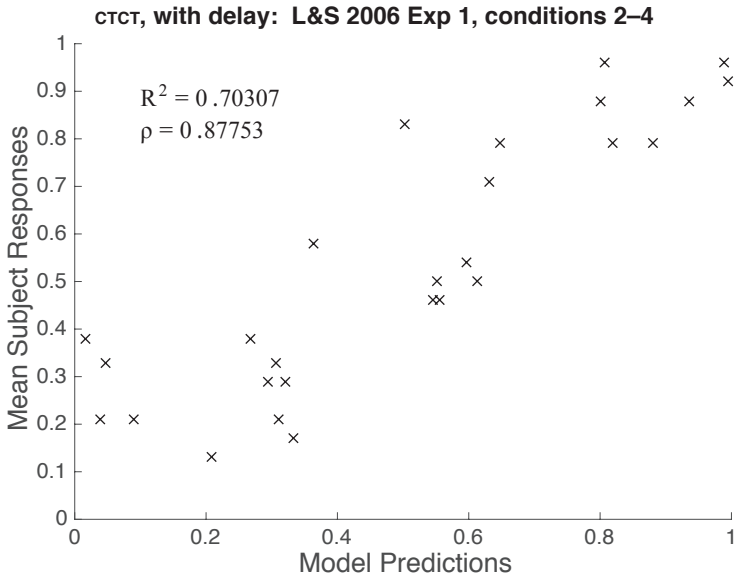


Figure 3.13: Results from our continuous-time causal theory model for Experiment 1 of Lagnado and Sloman[6], with decay rate metaparameter $\theta_d = 10^{-2}$, 100 loops, background rate $\theta_b = 10^{-3}$ and sparsity $\theta_s = 0.20$.

I find the same pattern where the model performs much worse on condition 3 than the others, with the same monotonic relation between condition 3's R^2 value and the sparsity value.

3.12.9.1 *Comparison with LBL[3]*

The continuous-time causal theory model outperforms the Wellen and Danks LPL Model overall in both the R^2 and ρ metrics (LPL: $R^2 = 0.47483$, $\rho = 0.7648$). This holds for both the case in which there is a delay and the case in which there is no delay. However, the LPL model explains more variance in Condition 3 than the CTCT model does, see Table 3.7 for more detail. In all cases my model qualitatively outperforms the LPL model.

3.12.10 Discussion

I have shown how to build CTCT model that identifies hidden structure on the basis of the timing and occurrence of one-shot events across a series of trials. My model's predictions not only have excellent qualitative fit with human judgements, but have excellent quantitative fit with judgements. Both the qualitative and the quantitative fits outperform the other available model

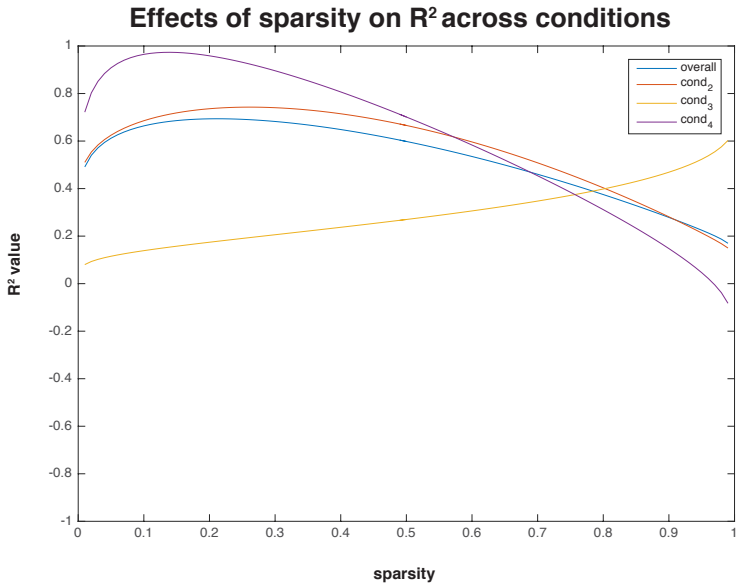


Figure 3.14: This demonstrates the idiosyncracy of condition 3's monotonic relation between the sparsity parameter and R^2 value. This suggests that in this case the errors between the models predictions and peoples' were due mostly to the model underestimating the prevalence of edges.

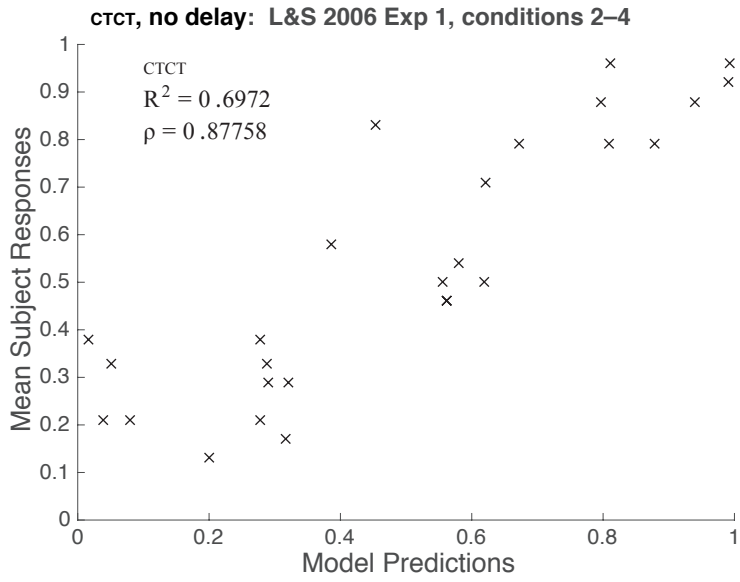


Figure 3.15: Results from our continuous-time causal theory model for Experiment 1 of Lagnado and Sloman [6], with 0 decay, 100 loops, background rate $\theta_b = 10^{-3}$ and sparsity $\theta_s = 0.21$.

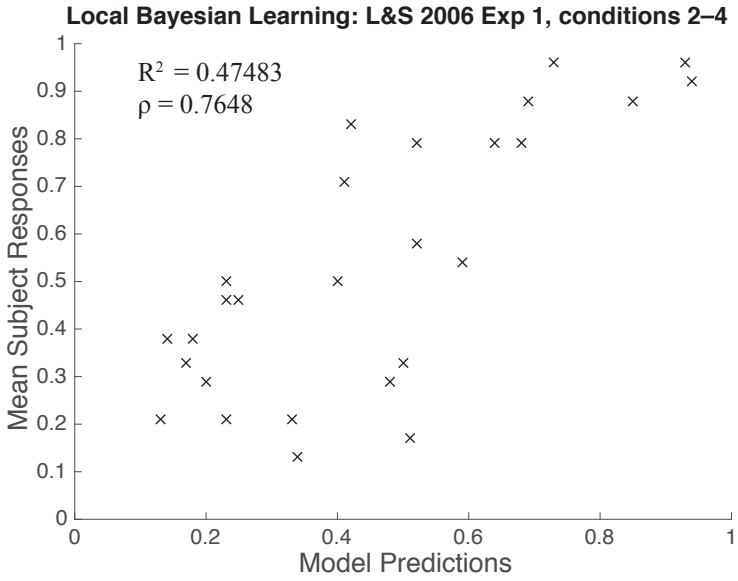


Figure 3.16: Results from Wellen and Danks[3] for Experiment 1 of Lagnado and Sloman[6].

Table 3.7: Table of R^2 (ρ) values including the CTCT results as well as those from Wellen and Danks'[3] LPL model. Largest R^2 values in **bold**.

	LPL	CTCT delay	CTCT no-delay
Condition 2	0.40 (0.71)	**0.75** (0.89)	0.73 (0.88)
Condition 3	**0.46** (0.83)	0.21 (0.91)	0.22 (0.92)
Condition 4	0.59 (0.81)	**0.94** (0.98)	**0.94** (0.98)
Overall	0.47 (0.76)	**0.70** (0.88)	**0.70** (0.88)

for this data – Wellen and Danks'[3] local prediction learning model – in general and on all but one condition of the experiment. One of the core theoretical distinctions between our accounts lies in Wellen and Danks[3] using a algorithmic-level analysis whereas I follow a computatational level analysis.

3.12.10.1 On algorithmic- versus computational-level analyses

The Wellen and Danks[3] algorithmic-level model contrasts with our approach at the computational level problem. I see the task as one of Bayesian structure induction without making specific commitments as to the manner by which the actual learning process occurs.

Sometimes the algorithmic level is called the representational level[⊛⊛]. This is problematic, and explaining why it is problematic helps bring into contrast the distinction between algorithmic- and computational-levels of analysis.

The Wellen and Danks [3] is committed to more than representational constraints, and representational constraints are embedded in every computational-level analysis as well. The entirety of this work on showing how representing temporal information as a stochastic point process embedded in a continuous subspace versus a discrete series of trials can be seen as an elaborate process for incorporating a representational constraint. Much of the flexibility of the framework I describe is largely the consequence of adhering to and operating over this representational constraint. Later, I will discuss other kinds of data that can exist embedded in a continuous subspace (states, values, rhythms) that are compatible with my framework, but cannot be accounted for without extensive formal retooling.

Nonetheless, I am not committed to any particular mechanism that fully describes learning causal relations in any particular individual. I merely expect that my models' average predictions will mirror those of the average person. I expect this on the basis of two assumptions: 1) my model and people are both solving approximately the same problem (which, in this case, can be formally encoded as structure induction over graphical models) and 2) that, however the solution is implemented, they are going to do a decent job at solving the problem. In doing this, I still have to be committed to the representation of our data (e.g., point processes), the data structures (e.g., directed graphs or distributions over directed graphs), and the inferential problem (e.g., identifying elemental links between nodes).

Algorithmic-level or process models as advocated for by Wellen and Danks [3] require you to state not only the representation of the data, data structures, and the inferential problem but also the exact manner in which all of the computations are going to occur. That is, they need to specify the exact internal representation that will be modified and what that modification will be under any circumstances.

3.12.10.2 Hunting down hidden causal structure

Mary Mallon left a wake in time and it was her undoing. The house that first drew attention to her had six cases of typhoid fever occur in the span of seven days, from which it was inferred that they were all infected at the same time being too close in proximity to be able to

[⊛⊛] I actively avoid reference to any process/algorithmic level explanation as a representational level explanation, and encourage others to do the same. Every level, even the computational level has some representational commitments, so that terminology is misleading at best.

infect each other. At first, she escaped notice as she had not long been an employee with the family. She had began working there as a cook less than four weeks prior to the first appearance of symptoms. She left shortly after the sickness took hold of the family; no new individuals were infected. These temporal coincidences were sufficient to motivate one epidemiologist to track her to obtain biological samples and, after being rebuffed in his attempt, to research other instances of sickness that had the same temporal pattern. In searching for these cases, he found that where a family moved out to the country, establishing a new home and soon succumbing to typhoid fever, there Mary Mallon was to be found. Where she had been employed at a home as a cook, those in the home became ill, and she left shortly thereafter. He decided he did not need the biological evidence because his "epidemiological evidence *proved* [Mary was a focus of typhoid germs]."(emphasis added) He continued his stalking, eventually leading to her arrest after which the sought-after biological proof was obtained. His actions may or may not have been just, but his inference was sound.

3.13 Inferring structure from continuous event streams

Hunting down causal mechanisms successfully can do a great deal of good for the world. However, there may be situations in which the data happen to be arranged such that many people are lead to believe in a false mechanism. Such faulty inferences are liable to cause great harm and suffering if they occur in matters of importance.

3.13.1 Vaccines, side-effects and inferring causal theories

Vaccines do not cause autism; no vaccines are causally related to autism spectrum disorders – either singly or in aggregation – and yet remarkable numbers of people believe (or fear) that they do. This specific link seem to have first arisen with regards to a specific instance of vaccination (measles, mumps and rubella) after the infamously retracted Wakefield et al. [221] report was published. A number of measles epidemics have blossomed in the United States and in the United Kingdom as a consequence of parents' empirically unfounded fears. Much of the literature that has arisen around this issue focuses on people's irrationality, the inability of physicians to communicate effectively, or the failure of methods to use emotionally evocative rhetoric so as to spurn change.

But chastising parents for their decisions or otherwise looking to explain their behaviour with appeals to failed decision making misses an opportunity to better understand the human mind from these cases. If we suppose, instead, that the nature of the problem tends to lead otherwise reasonable people to these beliefs, it is reasonable to think there might be features of the computational level inference problem that give rise to these phenomena. We will still appeal to aspects of human cognition, but it is in the spirit of understanding it in terms its attempts to succeed at causal theory induction not its inherent frailties. By examining the problem in this way, we reveal aspects of people's causal theories of disease, vaccination and immune systems. Furthermore, we when we cease looking at humans through the dusk-tinted glasses of the heuristics and biases account of cognition, it becomes clear that some of what are called wayward inferences may arise from features of the actual causal systems at play as well as the nature of the data itself.

3.13.1.1 Aspects of human cognition and causal theories

Searching for a causal theory. I have supposed throughout this chapter that humanity searches constantly for causal theories by which they can explain their observed data. Whatever it is that we mean by explanation, the absence of a causal theory that explains the origins of autistic spectrum disorders leads people to search for causal theories. As paraphrased by Eggertson [222] ,"The 'conspiracy theory' that vaccine manufacturers are hiding the truth about MMR and [autism spectrum disorders] is fuelled by parents' need to know what is causing [autism spectrum disorders]". This drive for causal knowledge is familiar, and there is little reason to presume that it differs from that which drives scientists, children or everyday people to search for causal knowledge and explanations.

Ontology: known causal entities. Vaccines are already candidate causes as they are known to be causally efficacious; we would not use them if they did not actually prevent diseases. And parents have valid evidence that their children are almost always harmed by vaccines. Puncturing skin is harmful in general and this holds for injections associated with vaccination — there is often redness, swelling and pain at the location that was injected. Even if the vaccine is not injected, any intervention on a complex system(like the immune system) has a chance to induce side-effects (effects other than those for which the intervention was primarily intended), and accordingly there are (usually mild) side-effects associated with the body's response to the vaccination. Society *should* be worried if parents were *entirely* unconcerned

about their children experiencing genuinely harmful events, but we should be just as (if not *more*) concerned about the possibility of the recurrence of contagious diseases with known epidemic potential in a world without vaccines. Though vaccines are causes that can be ascribed to mildly harmful effects — they are far less harmful than the alternative of not vaccinating children. The children themselves are at greater risk, as is society as it loses its herd immunity. Nonetheless, having entities available in one's causal ontology beforehand, along with accurate temporally and spatially localised evidence of minor negative effects makes vaccination events easy targets for those searching for entities with potential causal powers to place in their causal theories.

Plausible relations: the role of mechanisms. Mechanistic theories may play a substantial role in determining the persistence of these false beliefs.

Contact, barriers, and transport. The most plausible mechanism for a causal relation historically has been physical contact between the cause and the effect[37,223]. The skin usually acts as a barrier between the internal body and the external world, making it causally efficacious. Thus, any breach in the skin bypasses this causal mechanism and introduce into the body substances that would not have otherwise crossed the barrier. The efficacy of this breaching relation is *why* we inject things and perform surgeries; the motility of the circulatory system allows injected substances to contact the whole body (unlike surgeries for which contact effects are intended to be local). And we have learned the lessons from Semmelweis, dangers of outside contact are why we take so many precautions to avoid infection whenever and wherever body breaching occurs (single-use needles, antiseptic washes[188], local sterilisation pre-injection). People are concerned about substances toxicity only once the substance is within the body — whether ingested[193], inhaled, injected, or even absorbed[224] any substance can participate in a plausible causal relation after once it is inside the body. Thus parents would be right to consider any injection (vaccine or otherwise) as a plausible causal relation between a medical procedure and a chronic health problem, at least in the sense that testing the eye's pupilary dialation reactions by intermittently flashing a light would not.

Causal false friends. It is notable that a key part of original arguments from the anti-vaccination movements (including those in the fraudulent paper that triggered the most recent vehemence) emphasise a plausible (but evidentially unsupported) mechanism by which their causal claims could be enacted. The specific version that Wakefield et al.[225] invented ("autistic enterocoli-

tis") is one instance of a more general class of "leaky-gut" hypotheses, specifically stating that the vaccination irritated the gastrointestinal lining increasing its permeability and reducing its effectiveness as a barrier and allowing (unspecified) "peptides" to leak into the blood stream, pass the blood brain barrier, and reduce the availability of (unspecified) peptidase thus disrupting the management of breaking down endogenous opioid peptide production, causing neural dysfunction ◈◈. It is a complicated, unsupported, underspecified mechanistic hypothesis, but it *emph* a mechanistic hypothesis. And, abnormal intestinal permeability (the formal term for "leaky gut" hypotheses) has gained attention as plausibly involved in the causal systems producing various gastrointestinal diseases such as Crohn's disease, intestinal ischæmia, and graft-versus-host disease[227]. So even though the "autistic enterocolitis"[225] hypothesis was often so simplified as to be indistinguishable from other variations of other "leaky gut" hypotheses, it may have gained plausibility beyond its contact causality basis by association with other better supported hypotheses. This is one of the problems with causal theories, mechanisms that are identical from a lay perspective may be valid in one system but not in another, lending plausibility to causal relations that invoke mechanisms with falsely ascribed empirical support.

3.13.1.2 Aspects of problems of absence and the structure of continuous-time coincidences

There is structure inherent to causal theory induction in these cases that leads to patterns of behaviour and thought that can easily be seen as (at least) potentially rational, but which is often explained with reference to cognitive biases and human irrationality. Viewing these issues through the lens of continuous-time causal theories focuses us on how the structure of problems and data embedded in continuous time can explain behaviour otherwise thought to be irrational when not viewed in the correct formal context. I will focus on aspects that arise from the intrinsic sparsity of continuous-time problems and data, aspects that arise due to the preventive functional form when it aims to prevent contagious causes effectively, and aspects that arise due to the metric structure of continuous-time as it relates to the challenge of interfering events and spurious relations.

Omission bias, background causes and causal attribution. One of the reasons cited for accounting for parents not vaccinating their children is the "omission bias" ("the tendency to favour omissions (such as letting someone die) over otherwise equivalent commissions

◈◈ Though it is unspecified, presumably this is referring to dipeptidyl peptidase, which has been found to not differ in children with and without autism[226].

(such as killing someone actively")[228]. It is in this vein that rational models are rejected on the grounds that the "do not take cognitive biases into account".[229] Ritov and Baron[228] call this asymmetry a bias on the grounds that people exhibit it in situations where "relevant differences between omissions and commissions seem to be absent. For example, For example, choices about euthanasia usually involve similar intentions whether the euthanasia is active (e.g., from a lethal drug) or passive (e.g., orders not to resuscitate)." They focus on differences relating to moral considerations (intentionality and knowledge) rather than the formal/computational structure of the problem which give rise to the same "bias" as part of a rational inference mechanism. "Choices" to intervene in these sorts of situations are choices about point interventions in continuous time. As is the case with any such point process, the points at which we intervene on life are sparse in the set of all the possible points in which intervention could conceivably have occurred. Because of this, most of the time, we are all omitting interventions, and isolating any one instance or particular point of omission that merits causal attribution would seem to be impossible. Events that occur due to this persistent state of "omission" are difficult to distinguish from events due to the background causes.

On the other hand, acts of commission by definition actively intervene on the world, change it from the causal system governed by background causes and thus has easily attributable effects. And there are differences that manifest in the temporal pattern of omission versus comission. If one were to, to use the Ritov and Baron[228] example of euthenasia, to administer a lethal drug the effect would be expected to occur with a vastly different temporal distribution than the decision to not resuscitate; presumably there would be much greater variation as to when someone would die from omission than were they to die from a lethal drug which we would expect to have much more immediate effects. Additionally, because omissive plans rely on the consistency and momentum of the background causal systems they require much less deliberation to follow through on. In at least that sense, most of the time it is more accurate to say that no decision either is or needs to be made. And that holds even for those with the strongest sympathies in terms of active intervention. Even parents most in favour of vaccination spend most of their time *not* vaccinating their children — vaccinations would be far less appealing solutions were they to do otherwise. The failure to notice this asymmetry stems from the tendency to treat time in a discrete fashion in which this asymmetry does not exist (or at least does not exist quite as inevitably), and the failure to notice the asymmetry is what leads to the attribution of bias to the human mind. But a continuous time construction of the problem demonstrates that this line of thinking is not a sign of biased cognition, but merely sensitive to the structure of the system in which the decisions of this sort are made. Thus rather than

deeming it the omission bias, it may be more appropriate to term the commission inference.

Prevention silences its own support. The causal-absence dual of the omission paradox that gives rise to the omission bias is the prevention paradox. Vaccines present the perfect instance of the problem that arises from the prevention paradox. If a vaccine works and is used widely, its effects are difficult if not impossible to observe. The most ideal end case for vaccines is a world where no one becomes sick due to the diseases the vaccines are designed to prevent. But, if we that is the case, then vaccines are agents of the demise of their own agency both inferentially and existentially.

Problems for inference. We have eradicated smallpox, and as a result a smallpox vaccine is entirely unable to demonstrate its efficacy with data based on the natural occurrence of the disease. Our knowledge of smallpox and the eradication's success is based entirely on our knowledge of previous generations' fate. Suppose the memories of smallpox and its eradication were lost. People would no reason to believe that smallpox exists or that it ever existed, and that would still be the case even if they were performing some sequence of actions that included administering a vaccine against smallpox. Even if an enterprising researcher were to investigate that ritualised sequence of events, there would be no evidence demonstrating that the vaccine was a causally effective entity. Even if one had the hypothesis that such a thing as a vaccine might exist, one could undertake a microscopical examination of the vaccine substance and – if you did not know what you were looking for – nonetheless conclude that there was

By causal-absence dual I am suggesting (with analogy to mathematical duals) that we can emphasise at least two views on this kind of system that are related to the absence of events. The omission paradox attends to the fact that most of the time instances of events do not occur, but that does not merit inferring that there were causes or choices present that made them not occur. One could conceive of the possibility that a choice could be made or a cause could occur at any one of those points in time, but the default is for nothing to occur. This almost always merits a weaker inference from the absence of a choice/cause than from the presence of an choice/cause. The prevention paradox attends to the fact that most *kinds* of events do not occur ever, but that does not merit inferring that there were causes or choices present that made each potential instance of that kind of event not occur(which is why that event has never been observed to occur). One could conceive of the possibility that an event could occur that is of any number of kinds at any number of times, but the default is that most of those kinds of things never occur. This almost always merits a weaker inference from the non-occurrence of any instance of a kind of an event than from the occurrence of a kind of event. Both cases can be described as "most of the time events do not occur", but they emphasise different aspects of the problem; the omission paradox focuses on paucity of event instances whereas prevention focuses on the sparsity of event kinds. The omission paradox makes it possible to assign causal responsibility responsibly according to one's causal theories whereas the prevention paradox undermines those causal theories.

no "vaccine" in the sample observed. In the context of the 1854 cholera epidemic, Arthur Hill Hassall (who wrote the first book on microscopic anatomy in English) was specifically tasked by a government committee to study multiple water samples from across the London in order to identify whether there microscopic organisms in the samples that could be linked to the cholera epidemic. He finds examples of *vibriones* (a generic term for microscopic, motile elongated organisms) plentiful in the emissions of those infected with cholera and rarely otherwise, but concludes that they must thrive in the conditions produced in the body by whatever it was that did cause cholera.

Problems for existence. And, suppose that that world *actually* had no remnants of smallpox [◊◊], smallpox vaccines do not merely fail to demonstrate their efficacy, but they cease to be causes altogether. Prevention is a causal relation; in order for a preventer to be a preventer, there must be an event to be prevented. If there were no smallpox to have ever existed, there could never have been a smallpox vaccine. One could postulate a entity called smallpox as well as a hypothetical smallpox vaccine, but it is not clear what that claim would even mean, let alone how one would demonstrate such a claim without access to historical data demonstrating its existence (which by premise, cannot exist in a world in which smallpox does not and did not exist). We would look oddly at someone spouting the success of the campaign to vaccinate everyone against "snailpox" when there is no vaccine against "snailpox" and no disease to be prevented called "snailpox."

Prolonged preventions are particularly pernicious. Though I have been referring to the effects of preventers in terms of their efficacy, that can be interpreted in at least two ways. The first is whether the preventer works at all in preventing an event entirely (if it prevents a one-shot event like those studied in section 3.10 and section 3.12) or reducing the rate of events (if it prevents multi-shot events). The second notion of efficacy would be the duration of the preventer's effect. The longer this duration, the less evidence there can be in any finite period of time that the preventer has an effect. Even if the preventer's effect wears off over time, if it wears off slowly enough and occurs early enough in the history of observation it becomes difficult to detect the preventer's effect as being different from the background rate at which things occur. Events that are candidates for having been prevented would occur at times that are only

[◊◊] Though we have successfully eradicated smallpox in natural contexts, it is still studied in some laboratories often in the context of developing new, safer vaccines against smallpox. Thus virulent cultures of smallpox still live, keeping the possibility of a smallpox vaccine alive as well.

weakly related to the distance from the preventing event. The difficult of measuring such an effect is amplified for effect processes that only occur rarely. In fact, if one postulates the existence of a preventer with influence that *never* decays (it has a constant rate of cancelling events following its occurrence) the only opportunity one has to observe the preventer (in that process) as being identifiably different from the background causes will be the times before the preventing event occurred. If a preventing event of this sort occurred the moment that observation began, that preventer *is* effectively a part of the background process governing the effect process. The only hope one has of distinguishing the effect of the preventer then would be to view many samples of independent processes of the same kind where some of those processes did not have the event occur at the onset of intervention.

Useful preventative theories. And all of these features are exactly what we would want of our causal theories. It is not even clear what it would mean to presume the opposite. It is an understatement to call useless or ridiculous the notion that there are infinities of events not occurring that instantiate infinities of kinds of events that have never occurred because of an arbitrary number of preventers that we do believe to occur preventing the non-occurring events' occurrence possibly from the moment that the effect processes (that don't occur) would have otherwise been initiated. Such claims do not make sense and cannot relate data of the world. Our causal theories will invoke preventative causes whenever we can show that the events in question have a specific definition of what it would mean to have occurred and some evidence that such prevention actually occurred. That is most easily accomplished when one knows of and how to identify the entities and processes participating in the preventative relationship and when the prevention has a strong but decaying effect on effect processes that would otherwise occur frequently both within and across independent samples of those processes. Thus, we know vaccines prevent disease because we have observed the occurrence of those diseases historically and in populations not given the vaccine, and have observed their effects(the lack of disease) over time and across populations. The temporal aspect of observation means that data will only be available to be directly observed once, leaving later causal theorists to rely on that which has been recorded of the past to warrant their inferences.

For those who are not trained in the history of medicine or have not lived through the consequences of the disease any particular vaccine prevents, they will have little positive evidence for the effectiveness of the vaccine as preventing that vaccine. The more effective the vaccine is at preventing disease at all, the faster it is at ensuring that prevention and shortening the length of the disease, and the more widely it is used the less positive evidence there will be. Those

instances are where we will lean most heavenly on the structure and certainty embodied in our causal theories and the counterfactual and hypothetical inferences that they warrant. If someone refuses belief in a causal theory that relies on such a powerfully effective preventer, assuming they are reasonable (i.e., they are reasonably basing their inferences and judgements on the data they have access to and the causal theories they do believe), persuading to believe otherwise will be challenging if not impossible. For this reason, it is fortunate that we live in a world that suffers from the ignorance than the denial of the history of medicine. Vaccine reluctance/refusal seems driven more by a fear of potential negative side-effects than a rejection of the possibility of positive consequences on the basis of having not observed those positive consequences. However, it seems that the problems of distinguishing the effects of successful widespread preventative interventions (such as vaccines) from the background process generating the events will weaken people's beliefs about the magnitude of the positive consequences of of those interventions even if they acknowledge the existence of those positive consequences. Meaning that though vaccines' preventative status is not denied, the importance of those effects will be underestimated.

Temporal coincidences and small samples. It happens that the MMR vaccines are given at roughly the same time that the more prominent symptoms from ASDs manifest. This coincidence results in the appearance of a short delay between the event of vaccine application and the event of ASD symptom manifestations for some individuals. This precedence relation will occur for some children even if there is no overall trend toward precedence in either direction. The combination of that precedence relation with the short delay between the two events would by standard accounts be strong evidence in favour of a causal inference (descriptively, if not normatively). Most parents will have few children, and all parents will have any one child at most once, meaning that any inferences about the causal events and relations that have chronic effects on their child's life will have to be based on extremely small samples (often manifesting as cases of one-shot learning). In those small samples, it is reasonable to rely on the richness of the data that is available, and often that takes the form of continuous time delays. That would only amplify the strength of the evidence in favour of the causal relationship. Inference from commission would suggest that deviations from the normal state of affairs are candidates for causal events in a way that continued non-deviations are not, making active events like vaccinations even stronger candidates for causal events of any kind. Furthermore, parents spent most of their child's life not giving them vaccines, thus making the time when they do that much more distinctive a deviation. Parents existing at the same time will be us-

ing similar data and if that data is produced partially by the herd immunity existing because of widespread vaccination from early in childhood, they will have difficulty identifying the role and the importance that vaccinations play in keep their child safe from disease. Thus we are fortunate to live in a society where most people do trust in science's ability to successfully identify causal theories that successfully predict, intervene on and explain the world. If we did not, the structure inherent to the inferences, the problems and the data might make maintaining high levels of vaccination a much more precarious and volatile activity cycling between parents who do and do not believe in the efficacy and importance of vaccines based on what they are observing immediately around them. That we have such low levels of disease is a tribute to the power of scientifically based causal theories in governing people's decisions and behaviour.

3.13.2 Inferring structure in the face of real-time interference

Two of the most interesting aspects of continuous-time causal theoretic accounts of the false inferences regarding MMR vaccines and autism spectrum disorder are the ideas of multiple causal instances with aggregative effects and the problem of interfering events that obscure actual causal relations and suggest false causal relations. We are not here concerned with the functional form of the relation (vaccines generally prevent diseases and are not generally claimed even by supporters to prevent ASD), but rather whether different structures exist. These real-world examples illustrate a good deal about intuitive causal theories and their continuous-time aspects, but I cannot model those claims without amassing a great deal more data than would be possible.

We can learn similar lessons about these two particular issues by investigating Experiment 2 from Lagnado and Speekenbrink[7] using the same continuous time causal theory framework that I have used for other experiments. The study in Lagnado and Speekenbrink[7] was conducted to test the hypothesis that one reason that people could show a preference for short delays between causes and effects is because longer delays offer more opportunities for events to occur in the period of time between the cause and effect. To test this they designed a causal learning experiment in which participants viewed sequences of different kinds of cause events which may have been producing sequences effect events. The cause events and the effect events occurred multiple times over the course of the stimulus, which was presented in real time as a movie. In actuality, all the experimental conditions had the same causal structure (only one kind of event was related to the effect), and no stimuli were used more than once (each participant viewed a unique movie). The differences between conditions were based on the statistical

properties of the generating process: the delay between cause and effect events and how likely it was that an event occurred between the cause and effect (the probability of an interfering event, POIE$^{|||}$).

This manner of explaining a feature of human cognition (preference for short delays between causes and effects) in terms of the structure of the inference problem itself is well in-line with the rational, computational level project motivating CTCTs. Because the stimuli are indescribably temporal and rely heavily on the metric nature to obtain their structure, this is an excellent test case for the CTCT framework. Furthermore, though they can be grouped based on their generating parameters and used as repeated instances of the same condition(as done in the original analyses[7]), using each stimulus only once technically means that there are as many separate conditions as there are participants. Unlike traditional statistical methods which need to aggregate over many instances of the same condition, this kind of experimental structure does not pose a problem for analysis by CTCT. It can analyse each stimulus individually using common underlying parameters shared across all the stimuli to make predictions. And while we do not have data of the responses for each individual participant, we can aggregate the models' inferences across individuals to predict people's aggregate judgements.

In this case, I will infer structure much like in section 3.12. However, I will not be postulating hidden events, allowing a comparatively more straightforward probability model. For example, I will not need to rely on a detective probability model to approximate the generative model, I can work with the generative model directly. Additionally, unlike the Lagnado and Sloman[6] stimuli, Lagnado and Speekenbrink[7] use stimuli consisting of multiple causal events and multiple effect events not arranged in equal increments. This makes the LPL model[3] ill-formed (or at least underspecified) to be able to handle events of this variety. Indeed, the data are like that in Figure 3.12, but where we know the form of potential relations and instead are faced with identifying which of a set of causal relations hold between between multiple cause event sequences and a single effect sequence. To do this I use continuous-time causal theories to define an inference model over the set of possible graphs. Using this inference model, I obtain a posterior distribution over the graphs, and map functions of the inferred posterior distribution to different human judgements about the potential structures.

In this rest of this section I will model Experiment 2 from Lagnado and Speekenbrink[7] using the CTCT framework. I will first describe the formal structure of the experiments that will

$^{|||}$ Lagnado and Speekenbrink[7] called this the probability of an intervening event. Given my discussion about intervention in subsection 3.7.5 and the primacy of that notational convention, I modify the term so as not to cause confusion.

be needed to build models in my framework. Then, I describe how I implement the particular models in light of these formal properties, notably treating data from real-time event streams which I characterize as sequences of point events in a causal structure. I then use these models to compute the same inferences Lagnado and Speekenbrink[7] asked of their participants and compare my models' judgements to different human people's judgements which must be characterised by slightly different metrics. I discuss the fit of my models' predicted judgements, showing that they closely match average human responses in the conditions, suggesting that my framework succeeds both at characterizing the sequences of point events and at producing models capable of causal inferences and judgements that are comparable to that of human beings facing the same problems. Finally, I return to the discussion of how these sorts of issues manifest in the world before moving on to the chapter's general discussion in section 3.14.

3.13.3 Experiment Description

Experiment 2 of Lagnado and Speekenbrink[7] has the form of participants observing a continuous sequence of events (as a video) that represent the time-course of various kinds of seismic activity, specifically three kinds of seismic waves (which I shall refer to as A, B, C) and earthquakes (E). The goal of the participants was to infer which (if any) of the seismic waves were the cause of the earthquakes.

Earlier work suggests people will lessen their judgements of causal attribution between two variables if there is a longer(or more variable) delay between the occurrence of two events. However, this could either be because there is something specific about long delays between causes and effects, or that longer delays allow more opportunities during which other events could occur that are not causally related, thus weakening the connection between the original two variables of interest.** According to the design of the experiment – unknown to the participant – only one of the types of wave(A) was a cause of earthquakes, but sometimes non-causal waves would occur in the interval of time between the cause and its effects. This allows us to disentangle the two explanations for reduced causal strength due to longer delays. The length of time between the cause and its effect and the commonness of mid-interval events' sometimes were the primary differences between the experimental conditions.

** I should note that "more opportunities" is actually somewhat misleading as opportunities in plural form suggests that there would be a countable number of opportunities during which these events could intervene. It is more accurate to say that long delays allow for a larger, continuous amount of "opportunity"(a mass noun).

The experiment had a 2×2 structure. Delay-Length could be Long (mean delay between cause and effect $=$ 6s) or Short (mean delay between cause and effect $=$ 3s) — in both the standard-deviation is 0.1s. The probability a non-cause event occurred between the cause and the effect was Low ($\approx$ 35% of the time a non-cause event would occur between a cause and its effect) or High ($\approx$ 65%). These probabilities are approximate because the event sequences were randomly sampled and so cannot be expected to exactly match expected percentages. The authors chose delay distributions between the occurrence of cause and lure events to produce these probabilities in aggregate across samples.

They generated sixty datasets per condition that represented the time-stamps and identities of events that occurred in the movie. Of these, the first twenty datasets were used, with each of twenty participants participating in all four conditions exactly once. They were told that each animation would last no more than 10 minutes.✦✦

After each video participants were asked to provide judgements about the seismic waves that they had just observed. Participants were first asked to rate the extent to which each wave was a cause of earthquakes on a scale of "0 (does not cause the effect) to 10 (completely causes the effect)". This provides an "absolute" judgement of each wave's causal properties since the rating provided for one of the waves did not constrain the rating provided for the other waves. Participants were then asked for "comparative ratings, in which they divided 100 points amongst the three types of cause."

3.13.4 Building the Model

I treated the problem as one of structure induction. That is, given the knowledge that there are three possible cause variables ($\{A, B, C\}$) of the effect in question (E) and the data $\mathcal{D}$, I want to infer a posterior over the possible graphs linking the causes to the effects. Then I will use this posterior to compute measures analogous to those given by participants.

Data The data used to generate stimuli in Lagnado and Speekenbrink[7] are organized by the time-step (in milliseconds) that an event occurred and the identity of the kind of event (i.e., A, B, C or E). Though there were sixty generated sequences consistent with the design principles of their experiment, I used only the first twenty which corresponded with the conditions that they ran in their study.

✦✦ Though participants did see multiple conditions, I treat each trial independently rather than attempting to detect order effects. While I acknowledge its potential usefulness, addressing this is outside the scope of our analysis.

3.13.4.1 *Graphs and Parameterisation*

I considered graphs with any subset of three potential independent causes $\{A, B, C\}$. All causal links were generative and non-interacting with the other causal links. Thus the total rate of effects under a graph would be the superposition of all Poisson processes induced by the activity of cause-events according to the graph. As described in Pacer and Griffiths [156] this is the continuous-time analogue to the Noisy-OR parameterisation of a causal graph. Because causes were independent according to all graphs, the likelihood of their occurrences can be removed from our likelihood calculations.

In addition to a base-rate process $\text{PP}_{\lambda_\varnothing}$, which I assumed was a homogeneous Poisson process with rate $\lambda_\varnothing$, I allowed each cause-event ($t_d^{[X]}$) to initiate a NHPP with maximum rate (ψ_X) that decays exponentially(ϕ_X) relative to the the distance from the cause event($|t - t_d^{[X]}|$).

I sampled these parameters in a similar manner to Pacer and Griffiths [156]. I use uniform random variables ($u \sim U(-10^{\cdot 1}, 10^{\cdot 1})$) under a transformation ($\lambda_\varnothing = e^u$) to determine our initial timescale (in seconds), which acts as our base-rate PP. This creates a approximate scale-free baseline parameter ($\lambda_\varnothing \sim \frac{1}{\lambda_\varnothing}$) from which other parameters can be sampled. I sample $\psi_X \sim \Gamma(\lambda_\varnothing, 1)$ (the maximum rate induced by a single event of type X occurring) and $\phi_X \sim \Gamma(\lambda_\varnothing, 1)$ (the rate at which the intensity decays according to the distance in time from that instance) associated with each potential cause $X \in \{A, B, C\}$. Each cause instance($t_d^{[X]}$) produces a NHPP with rate function $\psi \exp(-\phi(t - t_d^{[X]}))$.

Because the baseline distribution is scale-free and defines other scales, these parameters are not "fit to the data". A "misfit" baseline scale produces overflow, underflow or other numerical and computational issues that result in model failure. But, any success can only stem from the model's structural commitments and the relation to the modelled data.

3.13.4.2 *Structure Inference*

For each graph ($G_\alpha \in \mathcal{G}$) and dataset (D) I take the sampled parameters ($\{\Theta\}_{m \in \{1,\ldots,M\}}$; for me, $M = 200000$) and compute:

$$\mathcal{L}(D|G_\alpha) \approx \frac{1}{M} \sum_{m=1}^{M} \exp(\ell(D|G_\alpha, \Theta_m))$$

I compute log-likelihoods ($\ell(\cdot)$) under G_α and Θ_m as follows. For computational efficiency, I eliminate events that do not alter other events (e.g., under graph $B \to E$, I consider

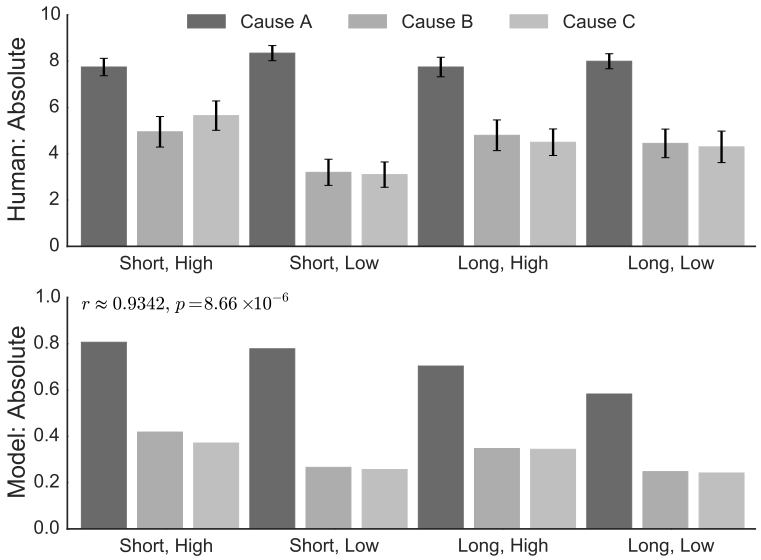


Figure 3.17: Top: Mean absolute judgements, Experiment 2 of Lagnado and Speekenbrink[7]. **Bottom**: m_{abs} model.

likelihoods of E- and B-events but eliminate A- and C-events).

Using the reduced event set ($\{0, t_1, \dots, t_i, \dots, t_n\}$) and the event-identity information ($\{0, X_1, \dots, X_i, \dots, X_n\}$), I can partition the observation period into a sequence of intervals ($\{(0, t_1], (t_1, t_2], \dots, (t_{n-1}, t_n]\}$). And then, considering each interval $\times$ event-identity pair ($\tau_{t_j, t_{j+1}} \times (X_j, X_{j+1}) \equiv \tau_j$) conditioned on previous events associated with the valid causes associated with the graph ($t_d^{[X]} \forall d \leq j, \forall X \text{s.t.} X \to E \in G_\alpha$) I can calculate the log-likelihood. The total log-likelihood:

$$\ell(D|G_\alpha, \Theta) = -\Lambda_{(t_0, t_n]} + \lambda_{\{t\}_0^n},$$

where $\Lambda_{(0,t_n]}$ is the log-likelihood component of the intervals,

$$\Lambda_{(0,t_n]} = \lambda_\varnothing (t_n - t_0) + \sum_{\tau_j \in D} \Big[\sum_{\substack{X, \\ X \to E \in G_\alpha}} \big[\tfrac{\psi_X}{\phi_X} (e^{-\phi_X t_j} - e^{-\phi_X t_{j+1}}) \sum_{t_d^{[X]} \leq t_j} \big[e^{\phi_X t_d^{[X]}} \big] \big] \Big], \tag{3.31}$$

and $\lambda_{\{t\}_0^n}$ is the log-likelihood component of the point events,

$$\lambda_{\{t\}_0^n} = \sum_{t_j \in t_j^{[E]}} \Big[\log(\lambda_\varnothing + \sum_{\substack{X, \\ X \to E \in G_\alpha}} [\psi_X (e^{-\phi_X t_j}) \sum_{t_d^{[X]} \leq t_j} [e^{\phi_X t_d^{[X]}}]]) \Big]. \tag{3.32}$$

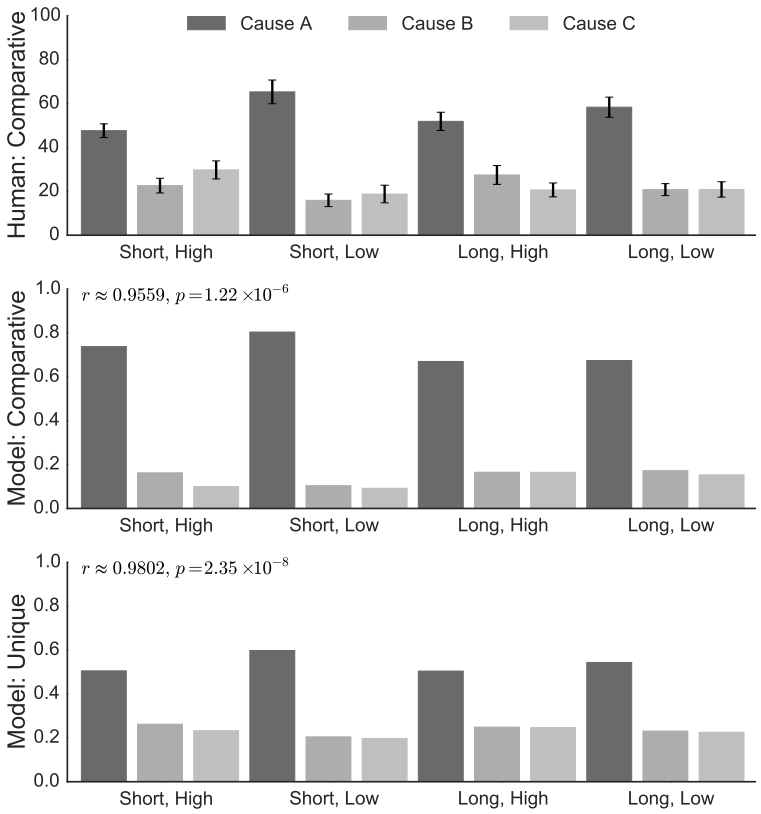


Figure 3.18: Top: Mean human comparative judgements from Lagnado and Speekenbrink[7]. **Middle**: m_{comp} model. **Bottom**: m_{unique} model.

Using the likelihood estimate($\mathcal{L}$) plus the prior for each graph (in my case, uniform over graphs $p(G_\alpha) \propto 1, \forall \alpha$), I can then compute the posteriors for all graphs,

$$p(G_\alpha|D) = \frac{\mathcal{L}(D|G_\alpha) \times p(G_\alpha)}{\sum_{G_\alpha \in \mathcal{G}} \Big(\mathcal{L}(D|G_\alpha) \times p(G_\alpha) \Big)}.$$

3.13.5 Comparison to Human Responses

People judged causes, not graphs; we need a way to map posterior probabilities $p(G_\alpha|D)$ to causal judgements. Lagnado and Speekenbrink[7] asked participants for absolute measures (assign each potential cause a value on scale from 0 to 11) and comparative measures (assign a total of 100 points to the three causes).

I will model judgements as statements about structure inferences (not strength estimations).

I interpret the absolute score in terms of a probability that a particular variable is thought to be present by marginalizing over the probabilities given to the graphs that include that variable is a cause. I.e.,

$$m_{\text{abs}}\Big(X \in N; p(G|D)\Big) = \sum_{G_\alpha : (X \to E) \in G_\alpha} p(G_\alpha|D)$$

Note if the complete graph were to receive all the probability measure, then $m_{\text{abs}}(A)$, $m_{\text{abs}}(B)$, and $m_{\text{abs}}(C)$ would each equal 1, and so their sum would equal 3. Thus, this is not a probability measure in the usual sense because I have not defined probabilities over causes, but over graphs.

However, by normalizing by the sum of these measures over all nodes, I can adapt the absolute measure to saying something about the comparative importance of the different nodes in producing the effect. This will sum to 1, but still should not be interpreted as anything like a direct probability of the cause being present.

$$m_{\text{comp}}\Big(X \in N; p(G|D)\Big) = \frac{\sum_{G_\alpha : (X \to E) \in G_\alpha} p(G_\alpha|D)}{\sum_{x \in N} \sum_{G_\alpha : (x \to E) \in G_\alpha} p(G_\alpha|D)}$$

Finally, we could consider the comparative prompt as implying that there is only one cause, and so we should only consider those graphs which attribute a single cause for producing the effect in question. In fact, we can say that under the restriction that only one cause may exist the graph including X as its sole cause *is* the measure of the comparative importance of X (since it is the only graph with that cause).

$$m_{\text{unique}}\Big(X \in N; p(G|D)\Big) = \frac{p(X \to E|D)}{\sum_{X \in \{A,B,C\}} p(X \to E|D)}$$

3.13.5.1 Results

I find an excellent fit between my models' predicted values and average human judgements for both absolute ($\rho \approx 0.93,\ p < 10^{-5}$, see Figure 3.17) and comparative judgements (m_{comp} : $\rho \approx 0.96,\ p < 10^{-5}$; m_{unique} : $\rho \approx 0.98,\ p < 10^{-7}$, see Figure 3.18) of the different kinds of waves' causal importance.

3.13.6 The danger of inaccurate causal theories and scientific reports

Unfortunately, the damage done by Wakefield et al.[221] will be hard to undo, despite extensive intervention. The causal claims have been refuted by many studies[230,231,232]. The paper was retracted[225]. There is widespread evidence that the study was fraudulent and that the primary author, Wakefield, had a conflict of interest[233,234]. But the Wakefield et al.[221] study has endangered many lives.

The article was amplified through the media and echoed in the error-checking-proof corridors of the internet, from where it entered people's minds[235]. Because of the ideas parents gleaned from these reports, vaccination rates decreased across the United States and the United Kingdom. That in turn has caused an increase of the rate of measles[236,237]. Given that we trust people's verbal statements of their motivations, we can trust that in this case we do know the causal mechanism at hand[235].

A scientific paper making a false claim about the cause of a disease has itself become a cause of disease.

3.13.6.1 Thiomerosal: correct ontologies, missing mechanisms, and unfortunate coincidences.

One of the other primary accusations of links between vaccination and autism illustrates a danger that can arise from inferences made with regards to causal theories that are mostly, but not entirely correct, especially when there is data that supports one line of action according to the incorrect theory and a different action according to the correct theory. *Thiomerosal* was a substance used as a preservative in a number of vaccines (though not the MMR vaccine) and the causally active component of thiomerosal was ethylmercury. Ethylmercury has antiseptic and anti-fungal properties. Contemporaneous with the Wakefield et al. fiasco was an overdue rise in awareness of environmental mercury toxicity[238]. Methylmercury and vaporised elemental mercury have severe acute and chronic effects as they stay within the human body over periods of time, and dimethylmercury is so toxic as to cause death within a year of being exposed to only a drop or two of it through protective gloves[224]. Ethylmercury appears to be comparatively safer, partially due to the relative rapidity with which the body eliminates it. But these distinctions are relatively nuanced, and the common causal theory has an ontology that includes only *mercury* with a poisoning causal relation to the human body. A campaign against Thiomerosal began and it was removed from most vaccinations by 2001.

Perhaps most ironic is that Thiomerosal is that it was included in vaccines specifically for the the safety consequences of doing so. In the early days of widespread vaccination, the most

dangerous aspect of vaccination was the possibility of bacterial contamination. In one of the surest cases of inoculation directly contributing to negative health effects (the 1928 Bundaberg incident) 12 of 21 children died within 25 hours of injection from a Staphylococcus infection[239] in the diphtheria vaccination mixture (those who did not die had abscesses at the injection site that teemed with Staphylococcus bacteria). What made bacterial contamination so dangerous was partially its hidden causal nature — in the 10 days prior to the Bundaberg incident 24 children had been injected with the same vaccination mixture and suffered no ill effect as it had not yet been infected by staphylococci. Given the pressing concern, great effort was put into finding appropriate anti-fungal and antiseptic preservatives. But many of the preservatives tried early on either weakened the effectiveness of the vaccine itself or were needed in such high dosages to be individually toxic to animal (and presumably human) subjects when injected subcutaneously. Thiomerosal was effective at low dosages (10000× lower than the concentrations needed by other existing preservatives) and had fewer negative side effects making it a attractive candidate for inclusion in a wide variety of vaccination. For most of its history the only challenges it ever faced were to its efficacy not its safety, though in large enough dosages (orders of magnitude greater than used in vaccines) it is toxic (but, note that the same can be said for water)[238]. Thiomerosal's vilification in the United States only occurred in full force in the aftermath of the general panic arising after Wakefield et al.'s work led to lay parents with intuitive causal theories of the cause of autism (self-labelled "Mercury Moms") organised to petition the government to ban thiomerosal from vaccinations[238]. Fortunately there were other preservatives by this point in time such that when the recommendation against Thiomerosal was initially passed in 1999, widespread vaccine contamination did not follow.

Human causal knowledge is only possible through the use of causal theories of the sort we have been discussing throughout this chapter. Causal theories and the inferences they comprise and allow are powerful influences on our decision making. But these theories are not infallible, and in fact can even be mostly correct(for example, missing nuance in how categories are subdivided and dosage thresholds are taken into account) while still leading one's beliefs and inferences astray.

3.14 General discussion

Our causal theories affect individual and collective decision making and are intricately linked to causal learning, reasoning, explanation, and judgement. Across contexts, including public policy and safety, causal theories will have intricate effects that to this point have mostly been

discussed vaguely using natural language or have been implicitly addressed using statistical methods that answer more limited problems than the theories actually pose and rely on simplified data structures (such as trials and assumptions of normality) to avoid worrying about the structure of the problem domain to focus instead on quantitative measures that can be trusted (due to the law of large numbers) to be amenable to the same kinds of statistical tests across domains. Human experience will never provide the number of samples that would be needed to support even a tiny portion of the current totality of human causal knowledge. Instead, we need to rely on the richness of the data in concordance with our theories about how the data are to be organised in order to make strong inferences. Human causal knowledge is richer than a flat aggregation of hypothesis tests, it has rich structure and is capable of dealing with complex data and inferences.

To fully comprehend the nature of the causal theories that play a role in human causal inference, we need to have a framework rich enough to express them, and I have made progress in the preceding sections. There, I have shown how to build formal models of causal theories within the CTCT framework. I have shown how to infer generative and preventative forms from rate data. I developed a integration approach to handle functional form inference for tabular data about one-shot events. I have shown how to infer distributions over a graph simplex to infer the form of relationships on data that occurred in real-time, where both causes and effects can occur multiple times. I have shown how to identify hidden mechanisms from one-shot occurrence data over repeated trials. I have shown how to extract different kinds of causal structure judgements on the basis of real-time data. I have done all this and demonstrated that my models' judgements accord well with human judgements about the same data (see Griffiths and Tenenbaum [30], Greville and Buehner [2], Lagnado and Sloman [6], Lagnado and Speekenbrink [7]).

In the large part, I have gone past theories of delays without much comment (see an excellent review in Krynski [163]). Theories of delays have hobbled progress in cognitive science's thinking about time and causal relations. Thinking exclusively in terms of delays is a thought pattern that only makes sense in the case where there is a one-to-one cause effect event mapping. But I have demonstrated (and modelled other experiments that demonstrate) that one cause can have many potential effects and that many causes can contribute to one effect ⇓⇓ and

⇓⇓ Some cases are indistinguishable in simple cases. For example, if you consider many causes contributing to one event versus one cause producing the event (but where the identity of the cause is unknown, but is the cause with probability equal to the proportion of the rate supplied) are mathematically indistinguishable in the case of structured point processes based on NHPPs like those I describe. These may be able to be disambiguated, but that will require further work and is likely to also require a

people will reason about them perfectly well. There is no need for the strict trial structures that make delays possible for causal inference to proceed, any suggestion that there was was due to the lack of expressibility in the mathematics used to analyse these experiments. At the same time, as shown in section 3.12, people *can* reason about one-shot events and causes (that may or may not have finite delays until their effects) that occur over trial structures. But we do not *need* to do so.

Though I have not emphasised it as strongly, the work extends the previous work on theory based causal induction in a number of other ways as well. Previously, the continuous-time work was limited to rates[30] and one-shot events that had unstructured hidden causes that occurred once[••] and observable cause/event entities whose relations were structured by space (see the discussion about exploding cans in minor section 3.3.4.2). In my work on real-time causes with functional forms, I had multiple events associated with the same entities occurring multiple times. In modelling, Lagnado and Speekenbrink[7] I not only address that problem, but I do so with multiple parents nodes as potential causes that are occurring in the same event stream. Surprisingly inference about multiple parents in this sense (where many events occurred each of which could be caused by any number of the parents as well as an unknown base-rate) had not been addressed by the previous work on theory based causal induction. My work on inferring hidden causal structure comes much closer to embodying mechanism inference than the previous work, as I model the case where the hidden events can play the role of the cause, but the effect as well. To be observed these hidden effects would need to be linked to some kind of observable event and inferred to have occurred at sometime before that observable event. To analyse this, I introduce the detective model of approximating the probability distribution, which bears a great resemblance to the kinds of mechanism discovery from medical history.

My exposition identified historical instances in which features of the rich causal knowledge and structure played a more-or-less analogous role in the context of medical decision making and inference. But I have gone further demonstrated that these models can use the same kinds of continuous-time data available to humans (possibly represented in a non-continuous-time format) to match human causal inferences. Part of the goal in including this was to demonstrate the utility of the history of science and medicine for cognitive science research.

This is only the beginning of the greater research programme. The search for a comprehen-

---

new formal framework in which to express the events.

[••] Though multiple hidden cause events were considered in Griffiths et al.[240], though they are a somewhat different case and could not cause each other.

sive framework for encoding human causal knowledge will need even richer representations, especially with regards to how causal theories are defined relative to time. Other models for temporal causal and statistical inference will be fruitful sources for further study either by providing new frameworks for modelling, reframing a sticky problem, or for providing inference methods that can apply more generally. We will need to consider other forms of causal relations (beyond generation and prevention). Going further in the study of representations and reasoning about hidden causal mechanisms will be crucial to account for how people go about interacting with a world in which have no choice but to rely on and reason about causal processes that they cannot and do not observe.

3.14.1 Other work on temporal causal inference with structured representations

Many other models have attempted to address problems closely related to those that we report here. Understanding them was crucial in developing the above work, and there are many more gains to be had from a deeper analysis of how they relate to and could interact with CTCTs.

3.14.1.1 Dynamic Bayesian Networks

Dynamic Bayesian networks "unroll" the network over time, where the edge structure creates links from the nodes at one time step to the nodes at later time-stepsDean and Kanazawa [241], Ghahramani [242]. This avoids simultaneous cycles by adhering to the DAG condition within each particular time-step. That allows treating the problem as one of a traditional Bayesian network with a DAG structure, where a node in the network represents an iterated set of nodes indexed at each time step in the rolled out network. This approach has been extended to cover many other cases including continuous-valued time-series data with interventions (with links to Granger causality) [179] or even non-parametric (infinite) models of entities/processes/events that can potentially be hidden [243].

A key problem with the DBN approach has to do with its discretising time. When applied to situations where there is a natural continuous-time metric, this approach requires discretising that metric, which introduces the problem finding the "correct" granularity for the time-steps. This granularity problem interacts with how one defines temporal relationships between parents and their children, which often are assumed to have effects only a small number of time-steps into the future. Indeed, Wingate et al. [243] claim something even stronger, saying: "sequences of high-dimensional observations are often generated as the result of latent events

in the external world, where events at time t interact to cause events at time t + 1." If we adhere to this version of the Markov condition for the world, defining the granularity becomes an extremely problematic issue.

In cases with long term dependencies, unrolling time may not only be impractical but it may be nonsensical. Continuous time causal processes of the sort we study have no problem with this; because there are always an infinite number of infinitesimal time steps that occur between any two non-simultaneous events, Note that without a notion of continuous time, it becomes a matter of definition that time steps are "equal". This places all causal processes that are supposed to follow the strictures of 1-step dependence to exist on the same granularity. If we start, instead with the notion of continuous time and integrate over it to recover discrete time, now the equality of discrete time steps is meaningful (it states that each integral over the time metric that makes up a time-step has the same value). We also could make it such that those time steps were unequal which (at least) allows distinguishing between the granularity of different processes. But then, if we not only need to bite the granularity bullet regarding the processes we observe, but also for those processes we do not observe. But *if* we already begin with the assumption that events are embedded in continuous-time — hidden processes will have events and causal relations that – if governed by genuinely continuous time processes like those we study – violate whatever time granularity is placed on the world due to the set of observed events. This is particularly trenchant for interval events which may be meaningless if they switch state in the middle of a time-step. For point events, two hidden events could occur at substantially different times but be encoded as simultaneous. Many events could occur, and they would be treated as one (if it uses an existence function) or would need to be represented as a count of events, not individually distinguishable events[⁂]

3.14.1.2 Continuous-time Markov processes

In order to understand some of the following examples, a brief introduction is needed regarding continuous-time Markov processes (see Ross [175] for a review). CTMPs can be seen as a Markov process defining the probability transitions between states and a Poisson process that defines when those transitions occur. You can see this as a trial based stochastic system where

[⁂] Note that if you were to treat them as individually distinguishable events, then you are effectively rejecting the grain that had been presumed as true for all causal processes. If one follows Della Rocca [244] and accepts that "there cannot be two or more indiscernible things with all the same parts in precisely the same place at the same time," then in order for these events to be distinguishable and of the same kind of process, they would need to not occupy the same point in time.

a change occurs on each trial, but where the trials last for metric, stochastic periods of time. For homogeneous CTMPs, these rates and probabilities can be decoupled, though they are usually presented as a intensity matrix that implicitly captures the information held in both of them. For the nonhomogeneous processes this can be a more complicated process. There is nothing in CTMPs that is causal; and while usually they could be given a structure/networked based interpretation, in the general case that viewpoint provides little value. However, they can be outfitted with structure and the relevant semantics to address these shortcomings.

3.14.1.3 Continuous-time Bayesian Networks

Continuous-time Bayesian Networks extend Bayesian networks to continuous time by using nodes with a finite set of states that are governed by continuous-time Markov processes for each node that are defined conditional to the value of their parent nodes[185,245,246,247]. That is, there is a separate CTMP for each value of the parent node, which are encapsulated by having (essentially) an intensity matrix that has the dimension of the product of the state spaces and only allowing transitions within between the virtual "states" that are associated with a particular set of parent values.

Originally these methods relied on particular functional forms for their transition times, specifically exponential wait-times which had the convenient memoryless property that allowed for easier interpretation. Attempts to generalise the wait-time distributions (for example by considering Erlang-Coxian transition times) introduced new problems (such as what to do about another variable which is in the middle of a transition when one of its parent variables transitions). I anticipate that many of these issues can be resolved, but I imagine success will be found by integrating richer), theory based knowledge into the models rather than using a generic modelling template. Already there have been promising results in applying CTBNs to cardiogenic heart failure[248].

This contrasts with my approach not only in that we allow point events to act as instant causal influences, but also because we can distinguish between stative and aggregative interval causal influences. One could imagine changing the rate at which state changed (the Poisson process) part of the CTBN without altering the Markov process, which would be comprehensible using the sort of framework we have described. It is much more difficult to see how one would have a systematic manner of altering the Markov chain; to have point events with effects on transition probabilities may involve manipulating values in a unnormalised space with the appropriate time delayed effects (e.g., by altering the parameters of a Dirichlet distribution

from which the transition probabilities are sampled).

In general, manipulating unnormalised quantities as causal effects will not have the kinds of straightforward interpretations for prevention and generation that we have been describing for point process intensities, because in order to be used they will eventually need to be normalised. This leads to the magnitude and shape of the effects due to any particular manipulation acting arithmetically on those parameters as being sensitive to where the state began prior to manipulation (see Aitchison [207] for more discussion on arithmetic operations over unnormalised simplices). That means that unless the state returns consistently to some kind of an equilibrium, no statement may be able to be made about the effect of a causal event. If it does return to an equilibrium state, then we may be able to describe effects in terms of perturbations from that state.

An equilibrium approach may be particularly fruitful in biological and other cases where feedback loops and regulators create equilibrium friendly conditions. Mooij and Heskes [249] found success in modelling flow cytometry data with causal cycles and equilibrium methods, but did so using structural causal models over real valued abundance measures of the different cellular substances. They were able to reconstruct causal knowledge in part because they were able to incorporate interventions. It is unclear if this kind of approach can be ported back to the CTBN problem, though I am hopeful.

To my knowledge there is no notion of intervention that has been proposed for CTBNs. However, Kan and Shelton [250] have applied the framework to Markov decision processes, which often involve making decisions in a way that can be described in terms of intervention over a state space, meaning that it is a feasible prospect.

CONTINUOUS TIME MARKOV NETWORKS Continuous time Markov networks (CTMCs) [251] are closely related to CTBNs, but focus on a slightly different modelling case. Rather than accounting for changes in the total state of a system, they want to model small deviations in features among a system that have global consequences — they use the modification of a genetic sequence encoding how a protein should be constructed as their motivating example.

3.14.1.4 Hawkes Processes and structured point processes

Hawkes [252],[253] processes are self-exciting point processes and mutually exciting point processes. This kind of self-exciting process is closely related to those that we describe, but when the relations between the different subprocesses have a richer structure than is required by general

Hawkes processes. These kinds of structured point processes are often (or, often can be) defined in terms of marked point processes, where each of the points has a set of marks (features) associated with it. For example, these can act as tags for tracking processes associated with nodes on a graph. If you join that with a semantics for the relations such that they respect the graph structure, you have imbued Hawkes processes with network that describes the dependencies in analogy to the graphs that we have been building before.

You can see a imbue Hawkes with a network structure in a variety of ways, though these approaches have not cast the problem as one of causal induction, introduced FPPs (as either a mathematical or a computational object[⊛⊛]), addressed the problem of intervention in continuous time, establishing various interpretations for the meaning of point events, or much of the other theoretical work completed herein.

Poisson Networks Poisson networks[186] are directed graphs where every node has an associated Poisson process whose rate is determined by the number of events from its parents that have arrived in a time window. This results in piecewise constant exponential wait-time distributions, but requires sampling schemes that are far more restricted than those we describe[186]. They also do not attempt to model the variety of phenomena that we cover.

Poisson Cascades One of the closest models to my own is Simma and Jordan[254] which relies on additive generative processes and multiplicative preventative(and generative) processes induced by events on marked point processes. It extends the earlier work done by Simma et al.[182] that focused only on the additive generative processes. The marks help define which of the events will induce which rate in which other marked processes. One can see these marks as embodying a graphical structure but they need not. They do not explicitly rely on logic based theories in the way that we do, and they do not consider the variety of problems that we do. Nonetheless the techniques used by and problems addressed by Poisson Cascades have a great deal to teach practitioners who wish to use CTCTs.

Bayesian Echo Chamber and reciprocating relationships Blundell et al.[255] studied how to model structured reciprocal relationships such as verbal turn taking and email networks using mutually exciting Hawkes processes. The different processes could take on differ-

[⊛⊛] FPPs as a mathematical object are interesting because of many formal operations and alternate viewpoints on their composition that they're generality makes possible. They are interesting as a computational object because of their straightforward sampling procedures.

ent causal roles (such as 'sender' and 'receiver' in the context of emails). By constructing it on top of non-parameteric models (including the Infinite Relational ModelKemp et al. [256]), they avoid the problem of needing to specify the number of entities (or even kinds of entities) that they need to postulate. This is a fascinating line of work that focuses much more on the large scale structure of the ontology inference than we do (and focuses far less on particular causal relations).

This work has already inspired work that in turn could be the basis of many interesting experiments: e.g., the Bayesian Echo Chamber [257] which treats interaction events themselves as being the point events of interest. It would not be easy to define interaction events in the current semantics of CTCTs, as the points would need to be associated with both of the processes and be simultaneous, which my formulation has excluded as impossible. We would need instead to postulate a new cross product node that exists in the crossed spaces of events occurring on one and another process, and then define those events that need to be coöcuring be the points on the space induced by considering the joint values of these variables such that the time is identical. I think that this line of work is promising and hope to incorporate it into later work.

Network discovery using the Network Hawkes model Linderman and Adams [258] are the only other marked point process/Hawkes process based model built on causal structure that is treated as generated by a prior over graphs. The Network Hawkes model prior is defined over exchangeable graphs (i.e., all the graphs that result were you to swap labels between them) rather than graphs with extensive hidden structure in the label information. Especially given the inferential techniques Linderman and Adams [259] that this approach warrants, there is likely a great deal of synergy to be found between these approaches and those described here.

3.14.1.5 Temporal Logic and Mechanism Inference.

Kleinberg et al. [260] demonstrate the ability to automatically infer biological mechanisms from theory enriched time-course data. They use background theories to organise the available representations for the entities/processes producing the data, (logical) functional forms for describing the relationship between these entities and processes, and information theoretic models for inferring the existence of these relationships based on the underlying annotated data. These three aspects map nicely onto my abstract description of causal theories, though in other respects this model is wholly unrelated to those described above being based on temporal logi-

cal relations rather than point processes.

Kleinberg et al. [260] apply this to gene-expression data (where gene identities were defined relative to the rate at which different sets of oligonucleotides that will combine/hybridize with the gene) annotated with information from the "Gene ontology" Ashburner et al. [261] (which groups genes based on fulfilling similar functional roles) to identify the developmental processes involved in the blood stage of the malaria bacteria *Plasmodium falciparum*. Their definition of processes requires a division of the sample times into (potentially unequal) intervals (time windows [OO]) that simultaneously segments time and gives boundaries for clustering the gene expression data, where a good set of windows "[captures] intervals of concerted gene activity, in which genes are clustered [in terms of] co-expressed elements". With a set of time windows and clusters for representing gene-functional data they build a model of the temporal logical relations that hold between clusters within and across time windows. Their temporal logic model is based on Kripke structures, and allows them to infer the underlying structure of the mechanism using the coöcurence data on the expression of different genes.

I have difficulty identifying how to conjoin temporal logic of these sorts and the work reported herein. They do not seem to be conceptually opposed, but the mathematics of describing them simultaneously is difficult (to say the least).

3.14.2 Exploring the conceptual universe in time: representations and relations

There are many kinds of data representations that one could use to represent the same underlying phenomena. But *data representations do more than represent phenomena*, they commit you to a particular ontology for the things that you believe the data are "data of". Even if you only accept such an ontology provisionally for the purposes of analysis and convenience, as soon as that analysis is put into action you have not only provisionally committed to the ontology but have *functionally* done so as well. Thus understanding the features inherent to the representations themselves (as well as the inferences they warrant and the potential relations between them). This is in line with the Kemp [140] programme to build a compositional frame-

[OO] These windows relate to what we referred to earlier as a posteriori discretisation. While the data was originally represented according to a continuous time metric, the time windows are defined relative to the sampled data points. It is worth noting that the Bozdech et al. [262] transcription data involved actively synchronised, but the model Kleinberg et al. [260] describe should be able to generalise to cases where the samples were not synchronised. However, doing so may require inferring a way to align windows in order to generalise across different time-course samples.

work that specifies the fundamental theoretical primitives available out of which our concepts will be built. Kemp [140] suggests that there are objects, features, and relations and elaborates by showing that different category structures are available depending on whether features additive or substitutive. We suggest that temporal aspects need to be built into the primitives themselves; some kinds of objects, features and relations may be irreducibly temporal. If we wish to formally model the richness of the engine for building theories that is the human conceptual apparatus, we need to be able to represent in our formal frameworks the range of data and phenomena about which people can reason. The experiments and models reported above demonstrate that people can reason with a variety of kinds of temporal data, making it all the more pressing that we make explicit the way time – continuous-time in particular – can be incorporated into the elemental structure of the conceptual universe. The remainder of this section will address this explication and, when possible, suggest formal solutions that could make this explication even more precise.

In the above work, we have mostly considered points and point processes as the primary theoretical construct needed to understand continuous-time causal induction. However, we have implicitly relied on other formal structures. In addition to points there are (at least) states, rates, waits, and weights. We have implicitly been using states, rates, and waits throughout this article, but they have not been examined in relation to one another, it is useful to be explicit about these distinctions. For example, we have used the fact but have not explained that events that can only occur once (e.g., death) cannot really be said to *have* rates, though the certainly have waits and can be calculated using the arrival view of a Poisson process which is endowed with a rate parameter. On the other hand, waits are difficult to understand if one violates the orderliness property of a Poisson process, as the formalisation presumes that there will be no waits that will be exactly 0 and that at all times you will be able to say which of the points precedes which other. Furthermore, while one would not be completely wrong in saying that our thinning parameters and general decay distributions for our non-homogeneous intensity functions are weights (of a sort), doing so would miss the key point of the distinctions we are trying to draw.

States. States are variables that always take on some value from a countable, discrete set of potential values at all times, and are notably characterised by some degree of persistence in taking on these valuesDean and Kanazawa [241]. Points then can be thought of as binary states that by definition lack the persistence property; they take on a value of 1 instantaneously (as defined by the Dirac δ) and then return to 0. So points *can* be seen as states, but because they

have a number of unique values

Rates. As we have discussed rates are variables that describe the average frequency of occurrence with respect to a unit (usually of time) that can be described on average or instantaneously.

Waits. Waits are the periods of time between two non-simultaneous events in the total process describing the superposition of all the relevant point processes. In any orderly process, you should be able to recover the wait time between any two events by the summation of the wait times between each of the sequential pairs of events that exist between the two in question.

Weights. Weights are real numbered values that are able to be observed at all times or (as in the case of velocity) derivatives of other weights that can be observed at all times (presuming some degree of higher-order smoothness).

3.14.2.1 State → State: Continuous-time Bayesian Networks

Continuous-time Bayesian Networks extend Bayesian networks to continuous time by using nodes with a finite set of states that are governed by continuous-time Markov processes that are defined conditional to the value of their parent nodes. In contrast, we consider the point events as instant causal influences rather that node states as durative causal influences. Any continuous-time Markov process requires a non-zero base-rate Poisson process determining its change-points, because it could not be decomposed into a unique (non-singular) Markov chain determining its state transitions and a Poisson process determining when those transitions occur. To my knowledge there is no notion of intervention that has been proposed for ctbns.

3.14.2.2 State → Rate(Points): Markov-modulated Point processes

Markov modulated poisson processes (mmpp) are not usually thought of in causal or even graphical terms. Traditionally they are defined as doubly-stochastic Poisson Processes (i.e., Poisson Processes whose rates are also a stochastic process) whose rates are governed by a continuous-time Markov process. This is not an inaccurate description, but it ignores the potential causal interpretation that they could have were we to think of these in graphical terms where a variable modelled as a continuous-time state process (specifically, a continuous-time

Markov process) is the cause of (and thus determines the rate of) a variable modelled as a non-homogeneous Poisson process.

If we take this causal, graphical perspective it is clear that we used this kind of a causal relation implicitly in section 3.9 while treating the state cause as an intervention. That is, we did not take into account the probability of the state cause, but rather assumed it to be true for the during the experimental condition it was applied. If the state of the electric field were to be stochastically generated (keeping the assumption that its state would be independent of the points generated), we could still perform inference using MMPPs though we also would want to model the transition rates between the states. We could still introduce intervention in this system by rendering periods of time during which the state was intervened on independent of the other periods of time during which the state was not intervened on.

3.14.2.3 State → Wait(Point(s) + State): k-shot, cause canceling events

When you trigger a firework it may not (and, hopefully, does not) explode immediately. As you step away there will be some delay until the explosion event occurs. During the time before the explosion, the firework observably maintains the state of being lit or is presumed to do so despite being out of sight. Or if it does not maintain the state of being lit, it will assume to have fizzled out and will no longer be expected to explode. However, once it has exploded, we can hardly say that it continues to be "lit" in any traditional sense. It may even be that there will be multiple explosion events (i.e., k-shot causes), but once the first event occurs this occurrence cancels the original state that initiated the wait-time process.**

Interestingly, there are two ways to interpret a process like this graphically. In one case, it is as I describe, where A a continuous-time state process causes X a continuous-time point process and on the occasion of X's first arrival, it cancels A. So the model would look like $A \rightleftarrows X$. An alternative model would be to describe A as a node in the vein of continuous-time Bayesian networks which is defined in terms of the amount of time it is expected to stay in the state it is in. At the moment when it leaves that state, it also *instantaneously* triggers the X with the appropriate rate/wait-time function. This graph would be more along the lines of

** Note that this contrasts with the version of one-shot events described above where it is merely the case that we define the first event to be the last event, as in this case the cancellation event could be interfered with in a way that death (for example) could not. If the fuse and explosive powder were unknowingly isolated from one another(or if immediately after explosion more explosive powder was added to the system), it is possible that the first event would not "cancel" the second. Instead, the second would appear to be a second instantiation of the same kind of event.

$\circlearrowleft A \to X$.

However, the latter version of the graph either needs to be slightly reformulated or it violates the orderliness property that we have maintained. Now in the joint event space for state transitions in A (which can be modelled as points) and the point process X two events have occurred simultaneously. If we sacrifice that, we sacrifice the possibility of a unique order, meaning that waits may not always be well-defined between any two events[***].

3.14.2.4 Weights → Rates → Points: Gaussian Modulated Poisson Processes

One can see variations in weights (variables with continuous values) as inputs to the function determining the underlying rate of events in a Poisson process. One can see our manner of using superposition and thinning to modulate real valued weights for each process. This is certainly relevant to the use of Poisson processes in spatial statistics in the case of kriging[265,266,267]. Recent work by Lloyd et al. [268] uses of nonparametric Gaussian processes as the input to a Poisson process; approaches hold particular promise being compositionally compatible with the kind of approaches we have pursued.

3.14.2.5 Points → Weights

Point effects can affect real number values. In fact, you could interpret the way models in section 3.10, section 3.11, section 3.12, and section 3.13 as cases where point events were changing a real numbered value when they altered the rate at which points were being generated (even if the exact timings of those points could not be observed).

This is slightly different than having points affect weights per se, which in theory can be directly observed at any point of time. This contrasts with rates, which can only be observed in the sense that they generate other observable events. For example, neurosecretory nuclei in hypothalamus release growth hormone releasing hormone, which stimulates the anterior pituitary gland's somatotropic cells to generate(or release) individual instances of growth hormone into the bloodstream. This changes the (real-numbered) density of growth hormoneKato et al. [269].

Controlling "weights" in the sense of "observable" real numbers could include point pro-

[***] This problem is particularly worrisome if we wish to move to a non-universal time metric. If we incorporate the insights of relativity theory[263] we lose even the possibility of non-transitive simultaneity such that waits will be even less well defined if we lose the orderliness property[264].

cesses that control spatial real values[▽▽] and their derivatives. For one example, consider the velocity of a watergoing vessel as controlled by one person creating individual strokes on either side of the vessel (called single-oar sculling). The angle of the oar affects the vector of the resulting force at each point, and so know the total effect on the vector one needs to know a number of real valued features of each point. But those are just marks on a point process. Fundamentally each stroke is a continuous-time intervention that applies force that continuously changes the location function.

3.14.2.6 Queueing processes

I mentioned the manner in which the Table 3.1 fails to truly duplicate the particle detector experimental paradigm, in that there were a finite number of cases out of which events could have occurred rather merely counting the number of events that occurred over a period of time. This is not the only time this modelling difficulty arises — for example, if you were to want to model synaptic connections between neurons as point processes, you would need to consider it in terms of the rapidity of events occurring among a finite number of molecules moving in space. It may be that the tools of queuing theory[271] – out of which much of the research on Poisson processes and counting processes originates – will be needed considered to address problems like these.

3.14.2.7 Structured sequences of events: mechanisms and music

One line of work that would be quite promising would be to investigate continuous-time causal theories handling of structured sequences of events. By that I do not mean just that the causal theories make inferences about events that happen to occur in particular sequences(see also, work by Bramley et al. [201]); but, rather, causal theories that specify *sequences of events* as part of their ontology, plausible relations or functional forms. One simple case of this can be seen in the domain of the traditional metaphor for deterministic causation (other than billiard balls), that is, dominoes falling. The sequence in which dominoes fall could be made stochastic by varying their density, height or the distances between them but fundamentally each domino falling (wherever it may fall) is a one-shot point event that acts as a cause for other

[▽▽] One might see spatial point process models for kriging and other geophysical prediction models as tacitly embodying this perspective[270].

point events[##] where the actual times and kinds of events that occur in relation to one another is semantically meaningful rather than incidental. In section 3.12 I identify stochastic mechanisms in that we postulate and reason about different systems of generic "hidden nodes" that can act as both causes and effects. Reasoning about complete mechanisms may be amenable to the same kind of analysis but it also may require more extensive theoretical, formal and empirical advances. I discuss this in greater detail in subsection 3.14.4 and minor section 3.14.3.7.

3.14.2.8 Historical Remnant Processes

Some causal processes leave signals in their wake. Causal theories can take into account actual historical events and their consequences, but to do so they will need to specify the signatures by which the temporal processes they describe are to be measured. For example, the Partial Test Ban Treaty ceased further atmospheric nuclear tests after October 10, 1963. As a consequence the concentration of carbon 14 – a radioactive isotope of carbon produced when atmospheric carbon is in the vicinity of a nuclear explosion – has been steadily decreasing. Carbon 14 (i.e., carbon with 6 protons and 8 neutrons) is a molecule that decays according to a (homogeneous) Poisson process, and its relative concentration (in relation to carbon 12 and 13) can be used to date when various carboniferous material was formed. This occurs because plants fixing carbon isotopes in a rough proportion to the availability in the atmosphere[272] can be studied and the ratios of the different carbon isotopes measured.

Other historical processes leave similar remnants and these remnants to coëxist in the same theory should be in accordance with one another. Tree rings (slices of tree trunk with samples from the tree's growth across its lifetime) are used to estimate the climate's properties across history[273]. Tree rings and radiocarbon dating are used to calibrate each other (). Human hair inherits the composition of the bodies out of which it grows. This is why Nierenberg et al. [224] analysed the time-course of mercury concentration in various lengths of hair grown after exposure to dimethylmercury. Our bones will recover from breaks but the mending leaves traces of the fractures. That these processes leave remnants in the world allowed scientists to identify the body of Diane de Poitiers from a mass grave and provide further evidence (based on gold concentration) that she may have died of gold poisoning. However, the work by Charlier et al. [274] is also notable in that they use carbon dating, but dismiss the results when they disagree with their other findings. They use their theoretical explanation of the method's inaccu-

[##] Per usual, there are caveats and exceptions to this claim. For example a domino that fell onto a spring that launched it back into place could be seen as a non-one-shot event.

racy due to the use of bitumen (asphalt) during the embalming process. To go further back in time, we can use paleomagnetism effects to date and track the history of the Earth's crust and land formations. That is, when underwater volcanoes erupt, the molten mass they emit cools and is magnetised in the direction of the Earth's magnetic field at the time of cooling making the magnetisation of the rock layers indicative of their age[275].

3.14.3 Other forms for causal relations

There are a number of other ways that causes and effects can relate to one another, many of which are primarily revealed only once we take continuous time's unique properties into account. Here I will detail some of these that will need to be incorporated into the CTCT framework if it is to be able to eventually capture the full range of human causal knowledge. One of the side benefits of this is that it may reveal aspects of problems that had previously remained occluded (e.g., the distinction between causing and enabling).

3.14.3.1 Susceptibility

> *[N]o one cause can be efficient without the aptitude of the body; Or …in a pestilence all would die.*
>
> Galen (as translated in Greenwood[276])

In all of the above case, following Kant[277], Shultz[200], White[195], Cheng[81] we thought in terms of causal powers, that is a static property of the entity/process associated with the cause events and determines the course of the causal interaction. However, a central piece of any causal theory attempting to account for medical phenomena will need to find a place for susceptibility, that is the role of the properties of the entity being affected as well as the properties of the effector. This is why in the case of section 3.11 we were able to treat the different samples as being instances affected by the same underlying parameter. If susceptibility would have been taken into account we would need to sample parameters describing these relationships

There are straightforward ways of describing these properties that plug-in to the current framework. For example, we could introduce an additional filtering term to account for different entities' static[♭♭] innate immunity against infection by a disease i in the context of active

[♭♭] It is less clear what it would mean to introduce a time-varying susceptibility parameter that is distinguishable from background variation and associable with an individual entity. One could imagine

preventer j, giving a different λ_k with a different $\Rightarrow_k$ for different entities:

$$\lambda_k(t) = \Big(\lambda_0 + \Rightarrow_k \psi_i \int_{T' \in T_i(C_i=1)} \delta(t, T')dT'\Big) \\ (1 - \Rightarrow_j \int_{T' \in T_j(C_j=1)} \delta(t, T')dT'). \tag{3.33}$$

We can see this introduction formally as playing a role in the functional form; but susceptibility properties could also play a role in determining whether a cause has any effect whatsoever. This kind of thinking plays a role in the development of causal schemata as described by Kemp et al.[278]. We could even use susceptibility as a third way to describe a one-shot process (in a sense a part of our ontology) by requiring a susceptibility parameter to have to exist in order for another event to be possible (e.g., one must first be alive to be susceptible to death).

3.14.3.2 Counting, build-up, and thresholds

Neurons with continuous time inputs have the property of transitioning between phases when they fire an action potential. This type of mechanism is captured by the Hodgkin and Huxley[33] model of how neuron membranes conduct action potentials using ionic gradients modelled by various capacitors and conductors. However, their model was idealised based on differential equations rather than point processes. Nonetheless, one could imagine a counting process that takes triggers an event after a certain number of inputs are reached (e.g., a fixed ratio schedule in operant conditioning terms[40]). Alternatively, it could build-up a signal that gradually decays after each activation (in a way analogous excitatory synaptic activity). Regardless, usually once the threshold is met a one-shot event occurs, and often the counting or build-up will restart, needing to meet the threshold again before another event can occur.

3.14.3.3 Enablers

One of the standard semantic problems in causal inference has been distinguishing "causers" from "enablers". A number of solutions have been brought forth (e.g., Hilton and Slugoski[203],

an active immune system that itself was time varying, but then we have given process its own entity (the immune system) that has an effect on the larger system. If on the other hand it varied according to time but not according to any known causal process (as that would again deprive the entity as being the sole "possessor" of the susceptibility),

Cheng and Novick [279], Wolff and Song [280], Sloman et al. [281]) none of which have been fully satisfactory *** Continuous-time causal theories offer great promise in the project of distinguishing these features.

In particular they offer great promise because they do not need to reduce to the single word ENABLER and are more expressive than the parameter estimation [279], dynamical systems [280] and Bayesian networks [281] that have been used to approach this problem before. For example, we can distinguish a enabling event (opening a glass door someone does not run into it) from a stative enabling condition (the state of the glass door being transparent and not observed by the runner). In this frame we can have enablers as preventers of preventers without needing to state a "normal" or "abnormal" state (as in Hilton and Slugoski [203]). For example, in the Challenger disaster we see the low temperatures acting as a preventer of a preventative state — enabling oxygen to leak into the combustion chambers by stiffening the O-rings which prevented their expansion to seal the oxygen off from the combustion chamber. To explain why this problem was not detected and the launch not cancelled may require invoking a "normal" versus an "abnormal" state, but we do not need them to merely define the causal scenario using the notion of a preventer.

This also allows avoiding some of the traditional problems with conjoining mechanistic and statistical approaches to causality as regards enabling causes. An enabler can act to boost some variable or process's (average) value over a threshold needed to achieve an effect, but would have no effect on the scenario were the variable process absent or of a lesser value†††. For example, reuptake is the process by which a neurotransmitter is absorbed back into the cell that released it (rather than crossing the synapse and being absorbed by the postsynaptic cell).

*** Because the existence of the sequence of papers elaborating and responding to one another is enough to establish the lack of satisfactoriness, it bears mentioning why Sloman et al. [281] fails in its explicit goal. Specifically, the logical model that is used in Sloman et al. [281] distinguishes between CAUSE and ENABLE in the deterministic case but fails to do so in the indeterministic case. While in the deterministic case, the error term present in ENABLE but missing in CAUSE distinguishes the two, as soon as fundamental errors are allowed into the meaning of CAUSE (as is defined in the indeterministic case) there is no difference in the functional form of the two relationships. If they had said that there would be presumed parameter values (or priors) for CAUSE and ENABLE, they would not have lost their distinguishability. However, the point of the paper was largely to identify the functional *form* of these relations and deny the parameter value interpretation, meaning their analysis was unsatisfactory and did not distinguish these ideas in the more realistic indeterministic case.

††† Note, the discussion of presence and absence is possible without discussing "abnormal" or "normal" conditions because of the sparsity of point processes in continuous time. Previous attempts have had to rely on the "abnormal" and "normal" distinction because in discrete time the 0 or 1 of absence and presence were interpreted logically, and therefore symmetrically [203,279].

Selective serotonin reuptake inhibitors (SSRIs) block the reuptake of serotonin and thereby effectively amplify the signal by effectively increasing the intensity measure of the signal (or, more exactly, it increases the integral over the serotonin density in the synapse over time).

One could also consider enabling conditions of a hierarchical fashion that act to "enable" events or processes that otherwise could not be well-defined. For example, acquiring citizenship enables one's participation as a member of a jury and thereby influences the decision of some court in one or another direction. It is not that it enables any particular decision in a particular direction, or even any particular decision ignoring the value (since the particular case that could be had is not yet defined). It merely enables the possibility that the one-shot process of deciding a verdict as a response to the one-shot process of being appointed to be a member of a particular jury that was convened in response to the existence of the one-shot process of a particular legal case. This is true even for laws that have not yet been passed at the time of the enabler. To say that acquiring citizenship enabled any particular event as opposed to a class of potential events seems ill-founded.

AMPLIFIERS There are also causes that will only amplify a signal once it is there, with particular attention to continuous measures. This can be contrasted with ENABLE which as Sloman et al. [281] notes can be interpreted to act as a necessary condition, AMPLIFY cannot act as a necessary conditions. AMPLIFY needs its process-object to be able to be defined as existing independent of the amplifier, which means that they are not necessary for that something (even under fairly generous modal logical interpretations). Enzymes would be a paradigmatic example: they make easier biochemical reactions in which some substances change into other substances but they themselves are left unchanged. If one speaks of individual reactions, it would seem reasonable to state that enzymes enable reactions. However, if one speaks of continuous measures such as the rate of interactions amplify makes sense while enable ceases to do so: e.g., contrast "The enzyme enables the reaction." versus the "The enzyme amplifies the reaction."

Amplification also seems to apply when there are interactions between causes that would be individually sufficient to cause some effect but which together produce the effect to a greater degree than would be possible than if either were present alone. For example zebra fish will seek out warm areas of water to induce behavioural fever, one effect of this is that it amplifies their innate immune response to pathogens in their environment. Individually, the fever and the immune system would be able to fight the pathogen (more or less), together they do more than either could do on its own. In this case, it is less clear whether there will always be a direction, meaning that amplify could be a "loopy" relationship in and of itself.

3.14.3.4 Stabilisers

For real valued processes, one could invoke a stabiliser (either as a state or as a point cause) that affects the variance rather than the mean value of the effect process. This is the role that an immersion circulator plays in heating water for cooking; by circulating the water and heating it constantly, it ensures that the water everywhere is close to the desired temperature (rather than merely having an average temperature that varies widely between parts of the container). Alternatively, a stabiliser could maintain the mean in a situation where there would otherwise be a change in mean. For example, you can view piloting a kayak with a sequence of strokes alternating on the two sides as a stabiliser for the vector in which the kayak goes, with more strokes on one side or another allowing one to keep a path against a wind or current that would otherwise change the net direction of the vector.

Alternatively, you can see a stabiliser as reducing the variance around a wait-time. This results in an increase in temporal predictability, and data with that property have been shown to increase the strength of causal inferences. A process like this may be necessary in order for real-valued equilibria to be maintained in the long run in spite of any permutations[249].

3.14.3.5 Tolerance and sensitisation: higher-order effects

Particular entities persist, and the actual *processes* associated with those entities could change on the basis of past events. That is, it may not be that only the effects or realisation of the process change as the form of the the effect due to a particular cause persists and decays (as is the case in the discussion in subsection 3.7.4). Rather, it may be that the form of the relationship itself changes as a result of the interaction. I believe it is possible to accommodate this as causal theories already use abstraction and meta-representations, but having forms that themselves are effects may require care in any actual implementation.

As studied extensively by Rottman and colleagues[282,183,283], when one is dealing with causal relations that occur over the same individual at different time-steps, this can introduce the possibility of changes in the form of the causal relations present in the system. These are most easily seen in the context of continuous valued causes, where tolerance can be defined as an increased magnitude of some feature being needed in order to produce a cause with this effect growing stronger as time goes on. Sensitisation is like tolerance, but in the opposite direction — a decreased magnitude of some feature is needed to produce the effect in question. It is less clear how to think about these phenomena when discussing events as they occur in continuous time. That is, the way they are defined in[183] as being structured over a sequence of trials

for which the metric time was supposedly irrelevant. In continuous-time relations, the metric temporal distance between events would likely matter a great deal.

3.14.3.6 Restorers and Regulators

There are a number of processes that may best be described by alterations to the ontology rather than the functional form or structure of a theory. That will almost inevitably change the available functional forms and structures, but there may be ways of mapping them into well established mathematical structures.

Restorers. So far we have discussed processes with symmetric and asymmetric states in which non-existence is the basic state, and existence is the deviation from that. It may be that you wish to include another basic state as a feature in your system; restorers can maintain a basic state by altering the processes' rate or intensity. One example of this would be doors that automatically close: closed is their basic state, the mechanism by which closure occurs is the process whose rate of change in the door angle that best optimises the closing of the door without negative effects (e.g., slams).

Regulator. This is like the timing mechanisms embodied by pace-makers in hearts, though that is more akin to a regulator, as it would describe the state-value as a precisely controlled entity continuously over time that hits several different modes of operation (a directed version of a Lorentz attractor around noise in the different intervals making up the stages). This process is called APD adaptation [284] and ensures that various heart contractions occur at the correct sequence and in the correct timeframe. Heart attacks result from failures to maintain this rhythmic balance. However, regulators can also regulate ratios or speeds, in a manner similar to how legal speed limits that change over space (and sometimes over time) or plumbing systems control water pressure. Order does not matter for a regulator, though regulators are also compatible with structured sequences of events.

3.14.3.7 Forms for modifying structured sequences of events

Some causal forms are indistinguishable from others unless we are describing the specific case of event sequences with rich structural information embedded within them, such as music [‡‡‡].

[‡‡‡] I thank Nori Jacoby for encouraging me to think in these terms.

Because these are far away from the Poisson based cases we have been analysing I will only describe them briefly. However, a full theory of continuous time causal induction would have to take these kinds of events into account as well. Indeed, doing so may be the key to completing the unification of statistical and mechanistic causal models described in subsection 3.14.4.

Hasteners Some causes act to hasten the occurrence of other processes. This is not clearly different from causation by increased rate in either a one-shot case[§§§] or a case with multiple unstructured event sequences [¶¶¶] to the events. Accordingly, it may be difficult to detect hastening causes in such systems (see Lagnado and Speekenbrink[7] for an example of people not being as sensitive to hastening), once one accounts for the increased predictability that can occur because of hastening[164] it could produce some effect.

However, in the case of structured sequences of events (such as music) where events have an identifiable substructure, we can distinguish between hastening and increasing the rate of a randomly arriving event. Specifically if we were to apply an *accelerando* to a musical score, the musicians would increase the rate at which they would play their notes, but they would do so in a coördinated fashion. The overall structure would be maintained. If we were to merely increase the rate at which a random process occurred, the structure of the process would be reduced. Structured event sequences will often be defined in terms of wait-times, sustain times and a common time metric across many variables (e.g., multiple staves on a musical score). In contrast point processes are defined over a unstructured space(or only minimally structured) and so doubling the rate could decrease whatever structure would already be there. [***].

Desynchronisers and metric shifts (swing) That said, events that were in synchrony could also fall out of synchrony (in phase → out of phase). Such a change would devastate models based solely on order (e.g., Bramley et al.[201]) that do not consider metric information. Nonetheless, there is no reason to forbid this as potential causal effect. Indeed, if the patterns that we learn as regular are sensitive to any kind of causal structure, our perception of the regularity and irregularity of events may depend heavily on the causal inferences we make, as

[§§§] Both hastening and rate increasing result in a shorter time-delay until the next event.

[¶¶¶] hastening and rate increases both increase the expected number of events.

[***] The structure that we have been providing is induced by CTCTs' functional relations between event types. We could recreate something that would be akin to music. This would require a causal regulation to some system that achieves internal regularity by using something like as aggregation mechanisms of many small events that makes up the basic definition of the second[172].

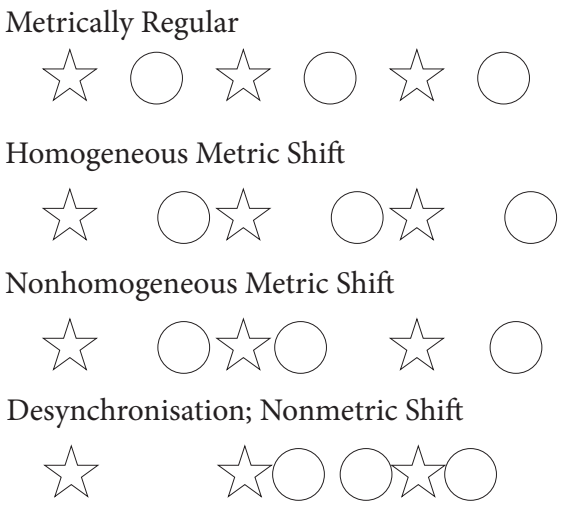


Figure 3.19: Illustration of metric shifts in a structured sequence of alternating events. The first case merely acts as a spatial "swing", the second case acts as a nonmonotone transform that does not affect the overall order of the events. The final case applies a transform to the second set circular events that turns out to be not monotone for the total set of events and eliminates the regularity that had once been there.

shown in Rhodes and Di Luca [285]. Similarly a global modulation of the relevant time metric in the vein of a "swung" rhythm ✠✠✠ would also be possible.

However these functional forms are going to be difficult to incorporate without the formal machinery to express them. Some method of describing a regular time metric in terms of basic stochastic processes would be necessary if that machinery is to be built of the CTCT primitives that are described here. Fortunately, based on analyses such as minor section 3.14.4.2 that may be quite possible. Describing transformations beyond globally monotone transforms that are able to apply only to some sets of events are going to be challenging to capture even in that framework (see Figure 3.19 for illustrations of cases where this may prove difficult). This is especially the case if the events become desynchronised and have their order perturbed.

✠✠✠ A swing rhythm is a rhythmic pattern that appears commonly in jazz. It most often is characterised by a shift in the way a sequence of 8th notes (which divide up a common time metric called the "measure" into 8 equal parts) would be played. Most often the first of a pair of eighth notes would be held longer, and the latter would be shorter (in order to fit within the same "measure"'s total measure).

3.14.4 Inferring statistical causal mechanisms

One of the most promising notes in Griffiths and Tenenbaum [1] is the idea that causal theories can act as a way to incorporate mechanistic information with statistical information. This chapter, particularly section 3.12, delivers on that promise.

We have presented a basis for inferring hidden causal mechanisms governed by stochastic processes using statistical information (or at least for inferring distributions over the space of possible causal mechanisms.). By that we mean, we have gone beyond simply inferring the existence of hidden causes on the basis of observed information. We have used timing and contingency information to identify hidden events that play the role of both cause and effect on each other so as to reconstruct people's beliefs about the structure of a causal system that they never directly observed.

We have identified on the basis of temporal events described in a variety of ways how we can use CTCTs to infer the form and structure of hidden causal relationships. If humans are using CTCTs, they are not reasoning about the aforementioned systems in terms of deterministic "billiard ball" mechanisms of the sort presumed by Descartes [223], Newton [286] and Hume [37]. People may be able to reason about mechanisms that are irreducibly stochastic but highly structured in their interactions.

Inferring mechanisms on this account can take on two flavours.

The first is one where nearly all if not all of the entities are known, at least potentially observable and linked to each other in a reliable and well identified way. This is the means by which we can investigate steam engines of the 1700s and infer the principle of their action by reasoning through the ways different pieces fit together and their compositions. We can do this even if we never see them actually in action, because we know the relevant properties of all the parts and what it will look like for them to act properly on each other. We can also infer in the other direction; having seen its behaviour and some of its physical features, we may be able to specify some of its internal causal structure (e.g., see Gopnik et al. [144]). These kinds of mechanisms can often be thought to be modality specific (e.g., Schulz and Gopnik [143], Schulz et al. [199]) and thus relates to the concerns raised by Shultz [200], Ahn et al. [287], Ahn and Kalish [288], White [195]. This too seems to be the kinds of notions that Kant [277] raised in discussing the inevitable propensity to view the events and processes of the world as possessing causal powers. [⊎⊎⊎] But only rarely is our knowledge of the constraints on causal systems so well specified as to

[⊎⊎⊎] However, Kant [277] also thought that Newtonian mechanics was the only way the mind could conceive of the world's dynamic processes and that Euclidean geometry was the only way that we could conceive of the world's spatial structure. So, perhaps we would do well to take his claims of *necessity*

make these determinate judgements.

The second flavour of inferring mechanism is well described by Griffiths and Tenenbaum [1], Keil [54] as the case where people have vague notions that some process is a cause of some other effect process, and that *some* mechanism brings about the process. That is, they presume that the relation they have identified has some unknown substructure in the causal mechanism by which the relation is manifested. They know this even though they do not know of what that substructure would consist. presume the existence of some series of causes linking the two, but do not know what that series of causes would consist of. For example, Darwin [289] provided an abstract mechanism by which, given some mechanism of heredity, he believed natural selection could proceed to produce speciation. He believed this even though the currently held "blending" theories of heredity (theories that Darwin endorsed) implied that differences from the mean would be swamped in the long run by regression to the mean [290] making phylogenetic speciation with distinctive features impossible(see West-Eberhard [11], Godfrey-Smith [291] for a more thorough introduction). One could imagine the second flavour as characterising an even richer class of mechanism inference. For example, the introduction of germ theory as a mechanism for disease etiology and communicability infers the existence of new *kinds* of causal relations and even higher-order mechanisms (such as the existence of animal carriers of the disease).

Capturing this entire range of phenomena so that we could begin to have a unified framework for describing various scientific advances throughout history is a challenging but fascinating challenge. This effort will be aided by describing some promising organising concepts that will likely relate to how this could be accomplished[⋉⋉⋉]. Two of these that seem most useful would be directly addressing nonparametric priors for the ontology, plausible relations, and functional forms (in the vein of those included in the causal theory work by Griffiths and Tenenbaum [1]) and making explicit the relation between macro-micro distinctions.

3.14.4.1 Nonparametric priors for kinds and numbers of processes and relations

When I have defined all of my processes, we have presumed the existence of directly detectable entity identities and known asymmetries between event types. This is not warranted if we want to be able to describe human causal inference in the real world. In the real-world inference problem, the number of entities is unknown, and perhaps even indeterminate since even the

somewhat lightly.

[⋉⋉⋉] This is in addition to attempting to incorporate other approaches to the problem such as that described in minor section 3.14.1.5

number of *kinds* of entities is unknown. Except by addressing this, how are we to be able to describe inferences of the sort in which John Snow postulates the existence of an unknown entity that transmits cholera by the fæcal-oral route thereby introducing both a new kind of thing and a new causal mechanism simultaneously[214].

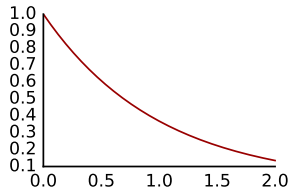
The challenge this introduces is held within size of the search space that suddenly becomes possible. Fortunately, inference methods from related models (see examples in minor section 3.14.1.4) should be able to be leveraged to make this more efficient. One advantage of these methods in particular is that they are specified in a non-parametric form already.

3.14.4.2 Micro-macro distinctions and gamma processes

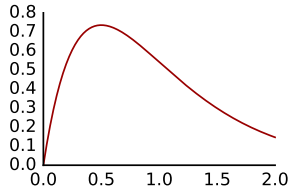
Quantum mechanics defines fundamental processes like radioactive decay to be irreducibly random (governed by Poisson processes); Newtonian deterministic mechanics has been rejected in favour of randomness[292,293]. In this sense, when our description of the world "bottoms out" in order to describe radiation events that govern the rate at which energy diffuses, we are left with Poisson processes. But the everyday world seems to have the deterministic structure Newton claimed it had. This is resolved by a general application of the law of large numbers; the aggregate predictions of many events at the quantum scale begin to appear deterministic when looked at from the macro scale. This is why our most precise clocks – those we use define the meaning of a one second[173] – maintain their regularity on the basis of counting many irregular occurrences of extremely rapid events⊛⊛⊛. The world we encounter works as Newton described because the overarching structure of many parts interacting with each other rapidly imposes order on the fundamentally disordered microprocesses.

One of the useful features of the Poisson process the exponential family form of its subcomponents. Though can be seen as the ultimate expression of randomness, eventually its activity becomes quite regular. As we've discussed, the wait time (in the homogeneous case) until the next event in a process is a exponential random variable. Note that the wait-time for the 2nd event to occur is just the wait-time of one exponential random variable added to the wait-time for a second exponential random variable (of the same mean). This is a Erlang dis-

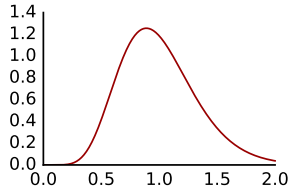
⊛⊛⊛ Until recently, our measurement standard and the most stable temporal sequence of events was the caesium based atomic clock which produced radiation events at the (definitional) rate of $9,192,631,770$ cycles per second (i.e., around 9×10^9 Hz)[172]. More recent optical atomic clocks are able to obtain even more rapid oscillations and potentially able to coördinate a single clock network throughout the world Komar et al.[294]. Nonetheless, these efforts are based on counting many events that are presumed to occur at irreducibly random intervals.



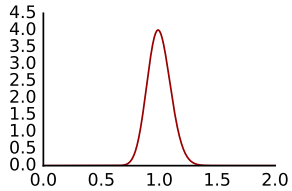
$\Gamma(\cdot)$ Time distribution $\mu_{\text{total}} = 1$,
with 1 $\text{Exp}(\mu_{\text{sub}} = 1)$ subcomponent



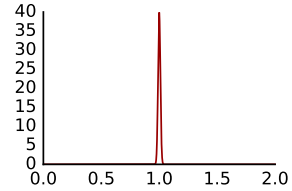
$\Gamma(\cdot)$ Time distribution $\mu_{\text{total}} = 1$,
with 2 $\text{Exp}(\mu_{\text{sub}} = \frac{1}{2})$ subcomponents



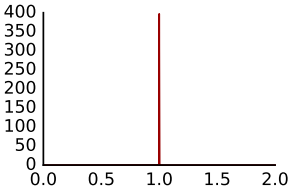
$\Gamma(\cdot)$ Time distribution $\mu_{\text{total}} = 1$,
with 9 $\text{Exp}(\mu_{\text{sub}} = \frac{1}{9})$ subcomponents



$\Gamma(\cdot)$ Time distribution $\mu_{\text{total}} = 1$,
with 100 $\text{Exp}(\mu_{\text{sub}} = \frac{1}{100})$ subcomponents



$\Gamma(\cdot)$ Time distribution $\mu_{\text{total}} = 1$,
with 10000 $\text{Exp}(\mu_{\text{sub}} = \frac{1}{10000})$ subcomponents



$\Gamma(\cdot)$ Time distribution $\mu_{\text{total}} = 1$,
with 1000000 $\text{Exp}(\mu_{\text{sub}} = \frac{1}{1000000})$ subcomponents

Figure 3.20: Illustration of the distributions of wait-times with mean μ into successively i.i.d. subparts with exponential distributions. The number of components ranges from $\{1, 2, 9, 100, 10^4, 10^6\}$. Notice how as the grain gets finer, the process ranges from perfectly random (1 exponential distribution with mean μ) to perfectly deterministic ($\lim_{n\to\infty} \sum_{i=1}^{n} X, X \sim \text{Exp}(\frac{1}{\mu})$).

tributed random variable, as is every kth occurrence of homogeneous Poisson process with rate λ ($F(X \leq x; \lambda, k) = 1 - \sum_{n=0}^{k-1} \frac{(\lambda x)^n}{n!} \exp(-\lambda x)$). This is a special case of the Gamma distribution which can be seen analogously as the sum of a real number l of Exponential random variables ($F(X \leq x; \lambda, l) = \frac{\int_0^{\lambda x} t^{l-1} e^- t \mathrm{d}t}{\Gamma(l)}$).

What is worth noting is that we have access to an argument like that used before in the discussion of approaching the definition of Poisson processes as a limit of Bernoulli processes as in subsection 3.4.3 and Figure 3.2 In that case, we found that as the number of trials in the same time period increased, the probability of an event occurring in any one trial gradually approached zero even as the rate was held constant. Here what we want to do is not to subdivide time into equal trials, but rather to subdivide a causal event into component sub-events to reproduce the macro-micro distinction described above.

Consider a one-shot process with a mean wait-time of μ and suppose its wait-time distribution is the exponential random variable with rate parameter $\frac{1}{\mu}$ (i.e., it is generated by an underlying homogeneous Poisson process with rate $\lambda = \mu$). Suppose instead that this was not generated by an atomic event, but a compound event composed of two subevents that occur in sequence and where each has a exponential random variable rate parameter $\frac{2}{\mu}$. The compound event will still occur with a mean wait-time of μ but its distribution will not be an exponential random variable but a Erlang random variable ($X \sim \text{Erlang}(\frac{\mu}{2}, 2)$ or $X \sim \Gamma(\frac{\mu}{2}, 2)$). We can further subdivide each of those unobserved subevents into two unobserved exponentially distributed subevents, making the macro-level X a composition of 4 i.i.d. with distribution $X \sim \Gamma(\frac{\mu}{4}, 4)$. In fact, we can divide X up into an arbitrary number of k unobservable subevents and its distribution will be $X \sim \Gamma(\frac{\mu}{k}, k)$, and that holds even if k is a *real* number of subevents (whatever that would mean). As k approaches ∞, we see that the wait-time distribution for X approaches the Dirac $\delta(x - \mu)$ where the event can be expected to be perfectly regular and occur exactly after μ seconds. In this way, we can range from complete randomness in terms of one uniform subevent to complete and exact determinism in terms of an infinity of uniform subevents (see Figure 3.20). Randomness and regularity are two ends of a commonly held spectrum of varied mechanistic granularity.

Much like we were able to model people's causal reasoning using Poisson processes – even of people who (likely) have no idea what a Poisson process is – people's mechanistic beliefs and inferential practices may be able to be described using common underlying formal toolkits. Even if noone actually thinks of deterministic mechanisms in this way , it could be that their

Though some may, consider the models of infinite springs as a mechanism for conveying waves on a string used in early work on analysis[295].

reasoning is well described by it. There may be other ways of examining people's beliefs that rely on stochastic information about hidden causal mechanism as well, such as the work by Buchanan and Sobel [296], Park and Sloman [297].

3.14.5 Feedback loops

Feedback loops are conceptually crucial. People often refer to causal feedback loops when describing complex causal systems (including biological systems [298], artifacts [298], and mental disorders [299]).

Feedback loops have continually posed challenges to formalising causal theories, and so they have been excluded from computational accounts of causal systems. Partially this was due to the difficulty in defining them in a formally coherent way without reference to time. For example, Pearl [15] had to introduce acyclicity as a premise of Bayesian networks in order to avoid paradoxes that arise when one considers synchronic cycles. To see why, consider that the definition of a directed edge means merely that the probability for a child node's values will depend on the realisation of the parent's values. If you then make the probability of the parent's values dependent upon the child node (i.e., creating a cycle or a loop), the child node is dependent on itself indirectly through the parent node.

3.14.5.1 Unfurling, not unfolding, time in feedback loops

Even if the world were divided up into discrete time steps, there is no reason to assume human cognition would be and plenty of reasons to assume it would not given the underlying representation. People do not have difficulties reasoning about arbitrary delays between events, but if discrete time is given a metric interpretation (i.e., a discrete time-step is given a particular metric value of, for example, 1 unit) this means that their causal relationships would need to be well defined over all countable infinity time-steps going forward, which could take on arbitrary shapes. This conflicts with claims from Kim et al. [300], who suggest that people only consider one time-step of a causal cycle (at least in the context of determining feature centrality in concepts).

But if people consider only one time step, how do you define that one time-step? Do we learn a completely different causal system for lights that take 1s to turn on after flipping a switch versus the same set of lights turning on after 5s? If they are the same system, then were you unable to interpret the difference between those cases in terms of their temporal relation?

In a sense these concerns are echoes the problem of dividing up time appropriately when moving from discrete to continuous time, but it is has some additional conceptual problems.

It is far from clear how to interpret this in the case that no event ever occurs (as in the experiments modelled in section 3.12). Does that trial last infinitely? In the case of a light switch failing to change the lights on one flip, is it a different kind of causal system if we flip the same switch again and the lights go on versus having to flip the switch twice before the lights go back on?[ʊʊʊ]

We have discussed one-shot events at length, and it is straightforward to see the "child" event in a one-shot system as occupying the "next" time-step for that system. But how do we reason about all of the different kinds of systems simultaneously, and how do we integrate them with one another? Scientific experiments often set up initial conditions, introduce some subject and observe the results of the trial; this allows time-steps to be defined relationally. If time-steps are defined based on the occurrence of other events then in what space do those three events relate to one another? What sorts of causal relationships ensure that they maintain the necessary form so as to maintain coherence?

Rather than seeing time as unfolding into a number of discrete chunks into which events much happen to fit, it is better to see it as unfurling into a continuous expanse with events falling wherever they may. We may construct the world such that when this unfurling occurs, we can identify clean boundaries around which to organise events. That allows taking into account experimental designs that are built to continually a trial structure for parsing the events that occur because of it. When you unfurl time, there is no "next" time-step even when there is a "next" event, and so we do not need to contend with the difficulties that arise from Kim et al. [300]'s claim that people only consider one time-step forward, because there is no such time-step for people to consider. The claim becomes meaningless.

3.14.5.2 Formally coherent notions of final causes and equifinality

One of the advantages of CTCT feedback loops is that they give a formal framework within which final causes or teleological explanations have coherent meanings. Specifically, if the process in question is to merit a final cause or a teleological explanation that means the process

[ʊʊʊ] Note, the notion of changes of state as opposed to state values as being capable of introducing further functional complexity into the domain of causal theories can be observed in work by Rottman and Ahn [301]. While they interpret their results in terms of grouping the events in some or another way, an equally valid manner of interpreting their results is in terms of state changes rather than groups around series of state values.

has a feedback loop which is organised to consistently move toward equilibrium states. This is best understood in terms of the outcome stability exhibited in equifinal systems.

Equifinality (stability of outcome state regardless of the route taken to reach that outcome), is a feature of describing systems with states that the process is "aiming at". Romeo and Juliet's dogged pursuit of each other despite many obstacles is often given as examples of an equifinal system. Thus, when Aristotle cites spiders' webs as an example of the purpose of a spider's existence he refers to the fact that under most circumstances a spider will produce a web of a certain kind and structure in the environment in which it finds itself. If the system is able to take its current state as an input to its own causal processes, no matter when or which events occur on that state, in a CTCT the rest of the causal system can be sensitive to those events and alter their relation to one another to bring about the "purpose" inherent in the overall structure of the system. From this perspective, a spider laying down silk first takes into account its representation of the current state of the overall web's structure and uses that to determine whether and where to lay down new silk. There are no preset periods of times at which these events of evaluation and action can occur for any arbitrary spider in any arbitrary system, but this feedback loop is continually available to the spider whenever it needs to be invoked.

Similarly, one can view the development of an organism from one cell into an adult kind/form (e.g., from egg to hen) as a process with developmental events that have as constantly present a cause that orients the process toward the "final" form (what we would today refer to as its genome) and a feedback loop from the current and local states to determine what events are to occur (and when they are to occur) to bring the entire process further along the path taken to achieve that final form♣♣♣. And, to be fully precise, we should think of that "final" form as a succession of locally final forms that lead to what we were referring to as the "final" forms. This may need to be introduced to account for the fact that there may not be monotonic development toward the final form as viewed in any particular light. For example, it is unreasonable to expect an embryo to begin growing a beak within seconds of beginning to divide into multiple

♣♣♣ It is in this context that theories of plasticity and development may begin to be formed as causal theories in the sense we have been describing. Without causal feedback loops that unfold over uncertain time intervals with causal precursor events rather than fixed time delays as the relevant features, it is uncertain how one could begin to formalise the remarkable body of work on developmental plasticity and evolution (see West-Eberhard [11] for a comprehensive overview of phenomena). West-Eberhard's *Developmental Plasticity and Evolution* is an excellent candidate for a source book for examples of how different kinds of hierarchically organised entities (phenotypes, genetic switches, behaviour, &c.) need to be considered in terms of their interactions and feedback loops in order to comprehend even some of the causal system (despite being able to isolate particular causal mechanisms with local, short-term effects).

cells. The governing cause that is directing the local path taken between the final forms will need to represent this time-course, but it will need to do so by relying on the inputs provided to it by the environment whenever it is that they are provided. There is no universal time-grid built into the genetic program governing the developmental process; the time-course relies on the timing causal inputs from the environment and the effects of the process itself. But since the development process itself may stochastic produce results, the system needs to be defined relative to the input events (external or internal) whenever it is that they occur.

The kind of procedural flexibility required for equifinal causal systems like these to work is not easy to describe using standard Bayesian networks without loops because their crucial feature is their self-influence. But it is difficult to see how such a flexible system could be expressed in discrete time, given that the input events cannot be guaranteed to occur in any particular order or at any particular times. It is likely that something more akin to a continuous-time causal theory that involves even greater abstraction than what I have included in my models. While I have stopped at generative models for producing continuous-time causal graphs, it is easy to imagine that the flexibility of these causal systems would require generative models of continuous-time causal theories, or generative models of generative models of continuous-time causal theories, and so forth. The keys are recognising that: 1) by looking at causal systems as existing in continuous time allows for encoding feedback loops in a coherent manner, and 2) higher-order generative models express the flexibility needed for those feedback loops to be appropriately equifinal by allowing the causes in the system to produce highly structured, regular responses to inputs that are stochastic and dynamic in their arrival, content, and composition.

3.15 Conclusion: The mind & time

If one were to take a single lesson from the work presented here, it is that time is crucial to our actual understanding of the world's causal structure. This should be unsurprising, the data we receive from the world is unavoidably and undeniably temporal. Even a minimally empiricist account of learning, reasoning or action would need to recognise that all of this cognitive activity would need to take temporal data as input. If we allow for self-reflection on these processes, that self-reflection will need to consider that these processes take time to occur.

That self-reflection reaches its pinnacle in the cognitive sciences, and accordingly time plays a large role in the methodologies of our formal studies of the mind and behaviour. That is time is integral to the psychological and cognitive sciences.

Theories of conditioning rely on time to define their subject of study. Classical conditioning is based on the temporal coincidence of an unconditioned and a conditioned stimulus[39]. Operant conditioning defines its reinforcement schedules in terms of rates and intervals, and it is unclear what extinction could mean if there were no notion of time[40]. Reinforcement learning theories are similarly tied up with time, often in terms of discounting expectations around future rewards[109,112].

Developmental psychology often needs to turn to indirect methods of inferring hidden mental mechanisms, so it is unsurprising that their inferences would rely on temporal information. Looking time paradigms are literally defined in terms of the duration and order of events and have been critical tools for the study of the development of higher order cognition[302,303]. The concepts studied themselves are intrinsically temporal notions, such as object permanence[304]. Even earlier methods unrelated to looking time attempted to find surprise and they too relied on time; specifically, they relied on the rate at which infants suck on a pacifier♦♦♦ as a mechanism for defining surprise[305]. The domain of infant and early childhood cognitive development itself – with its distinctions between the results of infants of various ages and the sequential development of cognitive abilities – would seem to be ill-defined without considering time.

Nearly every domain in psychology, when we reflect on it, has some methods intimately intertwined with temporal reasoning to hidden mechanisms of exactly the sort that we have described above. But this too should be unsurprising, for in Luce's[154] words: "Response time is psychology's ubiquitous dependent variable," because "one can infer back from the pattern of response times obtained under different experimental conditions to the structures involved." Perhaps the mechanism inference paradigm that we have described could be used to clarify just exactly what the formal structure of such inferences is.

3.15.1 Detecting mental activity as causal inference

And the inference to the mind on the basis of mental information is not merely a matter of speed, and it is not merely a problem for formal studies like those in cognitive science. Even infants seem to rely on temporal information to infer that something has a mind. Johnson et al.[306] have infants sit in front of novel oblong toys that that are able to beep and light up. Some of them had face-like features, others had none. Some of the toys would light up in response to (i.e., temporally contingent upon) infants actions (e.g., making noises and moving), whereas some of the toys would light up in response to another infants actions (i.e., it was an

♦♦♦ More accurately, a pacifier-like-object that may have been accompanied by a pacifier.

experiment with a yoked condition experimental design). Then the object moved forward and then turned left or right, and Johnson et al.[306] tracked whether infants followed the direction of that turn more or less often.

Gaze following (which turning in the direction that the object turned toward) is claimed to only occur with things that have gazes — that is, agents with the ability to perceive in the world. Many objects in the world move, but only those that have internal representations of the world will move so as to direct their attentional and perceptual apparatuses.

Johnson et al.[306] found that the infants followed the gaze in every case where the face was present; having a face meant having agency. But more interestingly, for my purposes, infants would also turn toward the object that behaved contingently with them even if it did not have a face. They did not turn toward the object that neither had a face nor behaved contingently. That is, the timing of the sounds and lights that acted contingently upon the infant's actions was enough to instill a belief that the object had mental events.

We can even model this as a causal inference problem using the CTCT framework, where the rate/timing of the reaction is increased by the occurrence of "infant action" events. In fact, we *have* modelled it by treating it as a structure inference problem using two continuous time causal graphs, where one graph has a generative relation (under G_1 the rate was increased by $\lambda > 0$) and the other has no effect (i.e., under G_0 the rate was not increased or $\lambda = 0$). We used the summary data that they included in the paper.♥♥♥ And like in section 3.9, to define the model we used an improper scale prior for the base-rate of activity ($\lambda_0 \sim \frac{1}{\lambda_0}$) and prior for the λ parameter based on the base-rate ($\lambda \sim \Gamma(1, \lambda_0)$). problem found results that accord with those found in their work. The log-likelihood ratio was heavily in favour of the generative graph in the contingent condition ($\log\left(\frac{\mathcal{L}(d_{\text{contingent}}|G_1)}{\mathcal{L}(d_{\text{contingent}}|G_0)}\right) = 11.24$) and was not in the non-contingent condition $\log\left(\frac{\mathcal{L}(d_{\text{non-contingent}}|G_1)}{\mathcal{L}(d_{\text{non-contingent}}|G_0)}\right) = -2.12$).

That is, there is at least a possibility that our manner of inferring agency can be interpreted as a causal induction problem of exactly the sort that we have been relating here. Which returns us to the beginning, to the inferential problem that Klausner faced. He heard a series of sounds with no apparent cause and so searched for the cause. The best evidence he found was that the sounds followed in close succession to another event: the snipping of the flowers. The scream happened only once and only briefly. When no snipping occurred there was silence. Unsurprisingly, beyond this general situation taken from the text there was no data included in Dahl's[103] story. However, that is enough to imagine that were the the events described to have

♥♥♥ Had I had access to the actual data this may have merited inclusion in the experimental sections.

occurred, the inference that Klausner made would be on similar grounds to those the infants implicitly used when they inferred that the toy had enough agency to be worth following its gaze.

I study human causal cognition because humans are by far the greatest instances of causal theory induction engines anyone knows of. The human manner of determining that other entities in the world are agents may just be one more instance of that causal inductive practice, and continuous time causal theories can mirror that inference. Perhaps, CTCTs could be the basis for an operational definition of what it is to have a mind from a generic human's perspective. If so, then as much as we trust other people to be succeed at identifying minds in the world, we might be able to build computational systems that we trust to do the same.

For what is time? Who can readily and briefly explain this? Who can even in thought comprehend it, so as to utter a word about it? But what in discourse do we mention to more familiarly and knowingly, than time? And we understand, when we speak of it; we understand, also, when we hear it spoken of by another. What then is time? If no one asks me, I know; if I wish to explain *to one that asketh, I know not.*

Augustine of Hippo [307], emphasis in original translation.

4

Conclusions on Explanation, Induction & Time

This chapter will lay out some of the issues that sit at the intersection of those I have been discussing but that have not gotten an appropriate amount of attention. In particular, I look at pairwise conjunction of Explanation, Induction and Time, and then consider issues that relate to all three simultaneously.

4.1 Explanation & Induction

Inference to the best explanation is crucial for understanding human inductive practices (see Lipton [308] for an wide-ranging discussion on this topic). Lipton's [308] strongest claim is that Induction simply *is* inference to the best explanation. The previous chapters shed light on this relation between explanation and induction in a number of ways.

Suppositions warranted by inference to the best explanation allow for "explanatory detours" that merit vertical inferences, in which a phenomenon is explained by reference to unobserved (and even unobservable) entities and processes. Having supposed those entities and processes to exist and to be embedded in a causal system that can account for the observations it is explaining, this will often warrant other predictions thereby guiding our information search.

That new information will provide a chance to either bolster or sap the proposed potential explanation's evidential support.

Lipton [308] argues that this framework allows us to explain features of theoretical induction and evidential support lacking in other accounts of explanation. Notably this includes explaining why a theory is more strongly supported by the successful predictions it makes (usually regarding data unknown at the time of the prediction) rather than by the data that the theory accommodates. This temporal assymetry has been proven difficult to explain by other accounts of explanation because they often rely on a logical support relationship for the two explanations.

Similarly, it allows avoiding Kuhn's [309] incommeasurability argument around crucial experiments: when two theories disagree about what even counts as evidence in an experiment designed to "discriminate" between the two theories, it seems impossible to describe a resolution to the disagreement. But inference to the best explanation allows describing both points of view, whatever the evidence happens to be according to one theory can be the evidence according to that theory, and likewise to the second theory. Then it is a matter of determining which of the explanatory theories are better. This is made more complicated by the introduction of unshared standards of what counts as a good explanation, but nonetheless, it removes the problem from its traditional frame as a logical impasse.

But even with these advantages there remain open questions of both normative and descriptive kinds.

4.1.1 Open questions about inference to the best explanation

In inference to the best explanation (generally) we accept the explanation that is best according some criterion, possibly the likeliest(or, more accurately, the a posteriori most probable) or the loveliest (that explanation that provides greatest potential understanding). But this leaves open at least three questions: why pick any explanation at all, why pick the best explanation, and why pick the best explanation according to any particular criteria. With this laid out normatively, there is the descriptive analog: *do* people pick an explanation, the best explanation, according to a single, particular criterion?

The work in the previous chapters addresses these questions, some more some less. It is easiest to begin with the descriptive questions and step back to analyse the normative ones in terms of a rational/computational level account of the problem.

4.1.2 Many explanatory criteria, matching and delayed decisions

Descriptively, from Chapter 1 we know that people's explanation generation and evaluation criteria are not well described by any one of the extant models of explanation choice from the artificial intelligence literature. Importantly, any model (even those we did not study) that would predict that single explanation would dominate all others according to all people will fail as well. There seems to be some degree of explanatory virtue pluralism at play at least across individuals if not within individuals.

Picking the best explanation is equivalent to a maximisation type decision rule (where the criterion in questions would be the utility function). However, people do not appear to be maximisers in all settings; there are some cases in which their decisions seem best modelled by probability matching (c.f., Luce's [310] choice axiom*) We took it as our null-model in Chapter 2 that people would match the posterior probabilities of the potential explanation assignments, not that they would maximise over the options. Given the apparently intrinsic rationality of maximising from traditional utility theory – even when applied to cognitive behaviours – if we are willing to consider non-maximisation decision rules for cognitive behaviours that leaves a large gap in our rational/computational-level analyes of these behaviours(see Eberhardt and Danks [75]).

The question of whether people *do* pick an explanation is actually underspecified. Certainly at times people pick an explanation, especially when prompted to do so, if they did not the methodology described in Chapter 2 could not have worked. Similarly, our finding in Figure 2.5 could not have occurred if people explain data immediately after observing data. Indeed, it is unclear how participants could have explained the particular datum that we asked them to explain before we gave them the case to be explained. This shows the two ways in which the question is underspecified.

Inference to the best explanation does not make an intrinsic distinction between the act of explaining a particular instance in terms of a theory enriched with a large quantity of other data or explaining the collection of data (rather than a particular instance). However our Figure 2.5 results cannot stem from an explanatory system that does not make this distinction. People could equally well have explained all the data immediately after viewing it, as there was no difference between the two conditions that differed along that dimension. The difference

* Luce's [310] choice axiom states that the probability of responding to one set of alternatives will be obtainable from the probabilities assigned to a larger, superset of alternatives by normalisation over the marginal probabilities of the items in the subset. This implicitly assumes a sort of probability matcing will occur.

arose in when we prompted explanation of a particular data point in relation to when we had people reconstruct their memory of the overall data they observed.

It is unclear when a cognitive agent will be driven to explain a general collection of data. It is equally unclear which data – without explicit prompting – will merit an immediate search for an explanation (possibly in terms of only the theory or possibly in terms of a theory enriched by other data in the theory's claims). Griffiths and Tenenbaum [311] provide an interesting account of coincidence detection in terms of theory-based causal structure inference analogous to our work in Chapter 3. In that they detecting coincidences as collections of evidence which are comparatively more likely under a novel causal hypothesis than under the given theory (usually a null hypothesis). This links their work to a causal version of the generalised Bayes factor that is used by the MRE model of generating and evaluating explanations. work may reveal a tight connection between the instances that prompt explaining and the

Inference to the best explanation is silent about *when* people will pick an explanation and whether it matters that they are explaining data (in general) or a particular case. Inference to the best explanation merely states that people will explain (at some point) and that, when they do, they will maximise a particular criterion that will privilege particular answers when making that choice. Despite these descriptive inadequacies, that does not mean that inference to the best explanation is not worthy of consideration. In fact, if we consider why people might violate these descriptive assumptions, it may reveal how we can salvage this account of human causal induction as inference to the best explanation by shifting the framework in which the problem is defined.

4.1.3 The Bayesian balance beam: the perils of choosing and maximising in the face of uncertainty

Inference to the best explanation assumes that we want to choose an explanation as part of our inferential and inductive processes. The most obvious alternative to this is to not choose one explanation but many. Holding many hypotheses about the state of the world is often called *beam search* in the machine learning literature. The extremity of this is to be irreducibly Bayesian in that you maintain a distribution over your beliefs about the different states that the world may be taking on.

If we were satisfied with the Bayesian solution, then we would not need to worry about choosing the best, or following any particular criterion. We could still get the positive effects of vertical inference; except now those inferences would need to be weighted according to

how probable each of the hypotheses was. Similarly the Kuhnian paradox can be avoided by marginalising over the underlying theories. It would just be that we use our distribution of beliefs as the input to our decisions.

But as noted, we *do* choose to explain, and we do reduce our hypothesis set to individual explanatory claims. There are resource-rational[36] accounts for explaining why we choose explanations — in resource rational accounts choosing to represent one (or a small number) of explanations in terms of needing to use limited mental resources most efficiently. However, it can be worth considering whether there are features of the environment or computational problem itself make such a strategy worthwhile. Doing so can show why the strategy of probability sampling is itself worthwhile, especially once we take into account that we will be choosing and the consequences of that choice.

4.1.3.1 Naïve realism and unique truth

A naïve realist view of the success of science and scientific explanation requires choosing explanations. That is, if scientists assume their aim is for their theories and the entities/processes they describe to capture the unique Truth of reality, a explanation choice model is the only way to succeed. If you only get "credit" for stating the Truth, and there is only one Truth, then the only hope of succeeding is to choose explanations.

4.1.3.2 Computational-level constraints and data preprocessing

Second, the Bayesian picture is built on a standard model of probability where there is a static state space where any observation can be made to accord with other non-mutually exclusive observations from the state space. But it is possible that such a space fails to describe the problem of human inference appropriately not only at the algorithmic level (or the resource-rational level) but at the computational level. For instance, one could imagine that the data input from our senses simply is of a form that cannot be directly conditioned on in the probability model defining the available potential theories.

Data may, in that picture need to be preprocessed in some manner before they can be incorporated into those models. For example, suppose the computational level problem faced by human causal inducers was limited to representations defined in terms of binary variables' occurrence over discrete trials. These are not representational constraints at the algorithmic level (as they are normally used), but at the computational level. Nothing is constraining the algorithms by which the problem is to be solved; only the shape of the problem itself is constrained.

But, if the world were to consist of continuously variable data, that data could only be evaluated by or incorporated into one of these theories theory once transformed into a binary, trial form. The decision would need to be made before the data could even be put in a form that the theory could reason about.

This precludes a solution in a Bayesian fashion if we adhere to the computational level representational constraint. If we were not constrained in our representational form for the elements of our theories, there would a Bayesian solution to this. That solution would involve computing joint inferences over the preprocessor decision making mechanism and the theories themselves. But, the decision mechanism for how to process the data cannot be a part of the same computational-level analysis. In order to handle the continuity of the world, the decision making mechanism would need to be able to take continuous inputs to transform them to the final discrete binary form. But that requires representational capacities outside confines available to these theories. Thus we cannot have a distribution over the preprocessor decision making mechanism, making marginalising over it (or performing any other standard probabilistic operations).

4.1.3.3 The dynamics of theories and data

A third way in which choosing explanations could be imposed in a computational-level analysis is by taking as a primitive that theories and data must by dynamically cultivated. In this view, the space in which theories can exist is itself is a dynamic entity. The data that is sampled gives the raw materials out of which new theories can be built. Those new theories in turn can merit new kinds of interventions, perhaps even providing the means of building new devices which in turn allow the collection of completely new kinds of data. The new theories that may need to be built to accommodate these new kinds of data may in turn reveal aspects of old theories that merit changing.

You cannot have Einstein's theory of special relativity without some notion like Maxwell's electromagnetic fields. And, you will not arrive at Maxwell's electromagnetic fields if your only exposure to electricity and magnetism are lightning and lodestones. It is not merely that a theorist who lacked access to precisely controlled circuits and magnetic generators would be dismissed by their peers. It is that it is not clear what it would mean for someone with no concept of generalised electric and magnetic force fields (of which lightning and lodestones are only two possible phenomenal manifestations) to propose that these fields were unified when considered dynamically. The data needed to build such a theory were only possible because of

the efforts put into developing apparatuses (e.g., batteries, wired circuits, and compasses) that could not only produce the phenomena, but could be seen to produce the phenomena in a reliable, systematic manner.

No matter what your computational-level analysis of learning, inference or induction, you will need to address the problem of sampling data from the world. In order to incorporate that data will require a sampling model (even if the result is to treat each data point equally). This does not stop the analysis from existing at the computational level. Assumptions like these are present in each of the analyses described in Chapter 3, which are paradigms of computational level analyses of causal induction tasks. But that is not addressing the problem of active sampling.

One wide-spread assumption[312] appears to be that once active sampling (rather than having data presented about which some sampling model needs to be assumed) is introduced the theory is then no longer able to be described at the computational level, but only the algorithmic level (or recently by resource-rational analysis[36]). This seems odd, as the "inactive" policy is available in which one actively samples to best match with the sampling model that appears to occur in the absence of active sampling. Perhaps the difference rests on the fact that even the unbiased active sampling relies on the existence of an explicit sample space based on your currently held beliefs about the "inactive" policy. But as long as the sample spaces of the actual "inactive" sampling distribution and the "inactive" policy are equivalent in the areas that they give nonzero probability to, the long-run behaviour of the inactive policy would converge to that expected from actual "inactive" sampling. In fact, in most real-world settings the "inactive" sampling distribution is the default for learning almost all processes at all times. There is always the policy of "do nothing and observe",† which would appear to be the kind of policy undertaken by participants in non-active sampling based learning experiments. Finally, regardless of the sampling policy, it may be possible to have a sampling model that takes into accounts any potential biases in the policy itself to bring it to accord closer with some model of randomly sampled data (in the vein of importance sampling). Which particular method one were to use might be the grounds for defining an algorithmic level analysis, but that would suggest that the feature that made the analysis algorithmic was the particular choice of how to

† This is not strictly true if one considers perceptual attention as a variety of sampling. However, if we do allow that as a primitive, then it is unclear how any phenomenon whatsoever can be treated as being anything other than active sampling. Because, then, in even experiments with no opportunity to engage in explicit active sampling in the experimental task, we must assume that the active sampling implicitly going on in perception accurately represents whatever distribution is present in the experimental task.

model the sampling problem not the simple fact that active sampling is involved at all.

If we take that distinction seriously, it means no studies of learning with first-person interventions are ever able to be analysed at the computational level. Furthermore, it relegates dynamic phenomena (like the sequences of interventions, active and inactive observations just discussed) as being unable to be accounted for by computational-level analyses. This would seem to unnecessarily impoverish the conceptual power of a computational-level analysis which is defined in terms of how one solves the problem optimally given its structure. There is nothing about that definition that states that the structure of the problem need not be a dynamic one. It would seem that part of this tendency results from the ease of analysis afforded by fully exchangeable probabilistic models which have largely been used in the context of computational level analyses[217]. Models like presume data can be analysed independent of the order in which it arrives. Rational process models (e.g., models using particle filters (sequential Monte Carlo), MCMC, importance sampling, and variational inference) have been proposed that take into account order effects and choices in sampling sequences in virtue of their ability to approximate these underlying exchangeable Bayesian computations[313].

Perhaps with a greater class of tractable models that are nonexchangeable and nonparametric (e.g., the phylogenetic Indian Buffet Process[314], the distance dependent Chinese Restaurant Process[219], the distance dependent infinite latent feature model[315], or the Bayesian Echo Chamber[257]) this tendency will be diminished. The hope in part is that in its unabashed nonexchangeability, the continuous time causal theory framework ‡ (CTCTs, as described in Chapter 3) can also aid in such a space.

Nonexchangeable, nonparametric priors allows the actual space of possible theories to grow in context dependent ways without needing to presume an algorithmic model. The space of theories will actually change depending on the data you happen to encounter. This leads to the conclusion that in these cases when you are sampling the world you are (possibly unwittingly) sampling from the space of potential theories that you will be able to entertain in order to explain the data that you gather. This completes the argument with which we began: in this

‡ To say that these models are nonexchangeable ignores the fact that some of their parts may be exchangeable even when the way they would deal with, for example, sequential data would not treat that data as exchangeable (or at least not in a straightforward way across all dimensions in question). For example, the Bayesian Echo Chamber uses a Dirichlet-Multinomial prior which is fundamentally exchangeable if one attends only to the role of the category identity. However the rest of the model describes the way these categories will interact with temporal dynamics, making the data points not exchangeable in the temporal dimension. CTCTs are also not exchangeable in some regards while exchangeable in others.

problem where the sampling procedure and the theory induction procedure are contemporaneous, you are indirectly sampling not only in the space of potential data, but in the space of potential theories.

4.1.3.4 Theory Gardening: sampling, pruning and inductive flexibility

Usually, if sampling theories or hypotheses is under discussion, the type of model in question is not considered a computational-level model, but an algorithmic one[316]. As a result, this dynamic perspective for a computational-level analysis is close to that of the resource-rational account of building theories. But they differ not only in emphasis but in their problem ontologies. The resource-rational view sees the computational-level account as being able to cover the space of every conceivable theory for all of the events that could ever occur simultaneously. It is, in a sense, "eternalist" — it presumes that the space in which the computational problem is defined is constant. For such an account any dynamic features of the problem arise from the fact that our resources are limited in our ability to explore the space.

While this approach is important, it seems to miss the point of the actual problem that people are solving in the exact case that I am trying to identify here. It dooms computational-level analyses to be forever disjointed from the problem that computational systems living in a dynamic environment must actually solve. That is, it may be that what is impossible one day is possible the next day or vice versa (no one alive in the twentieth century can aspire to be a professional dodo hunter). To deal with building a dynamic theory space to accommodate this dynamic causal environment in the case where we will be actively sampling data, explaining by choosing explanations begins to appear to be somewhat inevitable if only to be able to continue feeding data to and pruning the theory garden being cultivated in this process.

The real computational level analysis of causal induction will have to address the fact that human minds are often learning multiple theories at the same time. If we were to discuss this in terms of the lack of available computational resources or the number of samples for representing the different theories, then we would be working at an algorithmic or resource-rational analysis. But even from a computational-level perspective, this multi-task causal induction system faces constraints due to it solving many problems at the same time, if only because a finite amount of *data* can be sampled from the world. This means that certain aspects of some of those theories will develop more or less slowly than others. This holds even (especially) if the same data that is sampled bears on more than one theory.

And it is here that we can begin to see why a nonmaximisation approach would benefit an

explainer faced with this computational level problem once saddled with the consequences of explaining. Consider the most general version of our finding in section 2.4: that explaining can alter our memory of data to be more in line with the explanation than the data actually was. Suppose that you explain some data with respect to one theory, and thereby shift it toward your explanation — any other task that involves relying on that data will now be affected by your explanation. If you were to maximise immediately after each datapoint, and you reconstrued the data in terms of whatever the maximal explanation was (instead of probability matching) you would rapidly discover that you had taken away the nutritive data needed to maintain the full space of potential explanations and theories. Effectively, by consistently explaining and doing so with a maximisation policy, you begin to see the world through the lens of whichever theory/explanations happened to be favoured early on in the data collection procedure. Even if you pursue a disconfirmatory policy, the evidence you gather will be defined with respect to the theoretical entities/processes and concerns of those initially favoured theories and explanations.

However, that initial bias is not the core problem. That happens if you pursue probability matching as well. The difference is that if you were to probability match you have a chance of explaining with reference to theories and explanations that are not maximal but are good on other grounds, which in turn lead you to gather data that would have been inaccessible had you maximised the entire time. That is even with the constraint that the act of explaining (and thus data compression) is inevitable at the computational-level this strategy allows exploring the entire space of potential explanations, theories and data at least in expectation. A maximisation strategy can make no such guarantee. Resource rational and algorithmic/process level theories can produce similar results, Abbott et al.[317] provide suggestive along these lines in a simpler domain wherein particle filter resampling policies can induce primacy effects in a manner similar to the role we are giving explanation here.

Given that the probability matching explanation choice model will be unbiased in expectation, a system aiming to solve multiple tasks can rely on that data to be as it was originally observed (at least in expectation). A maximising system will ensure conflict between theories that gain support from different aspects of the same data. The consequence in that case may be that theories are treated as mutually exclusive when in actuality they are not. The totality of the mind will be weaker for this strategy.

This gives a solution to the problem raised by those arguing that maximisation should be the criterion of rationality if we are to be posing rational, computational-level analyses of cognitive behaviours like causal induction[75]. Though explanation and inference may have formal analo-

gies to decision making, with potential consequences for the memory for the data that support those "inferential decisions", they are not well characterised by unidimensional utility models. There are often many inferential processes proceeding simultaneously, and they may interfere with each other due to computational level constraints and the unavoidable effects of the "inferential decision making" process. In the case of inference in a dynamic world with multiple aims, maximisation will ensure inductive rigidity while probability matching will guarantee at least some degree of inductive flexibility. Nonetheless, it does so while also allowing us to make (non-inferential) decisions and to guide those decisions with inferences and explanations as they are needed. It just allows that guidance without handicapping its ability to guide the reasoner in later tasks (though one could imagine that frequently requiring explanations in the same vein would also lead to a stultifying effect on the available set of explanations).

4.2 Explanation & Time

Explanation and time are deeply related. We do not explain things from the future, and we would not be able to explain things quick enough to explain those things in the present — by the time the are explained, they have passed.[§] Thus we explain particular events in the past on the basis of other past events; this is the problem of causal ascription or the problem of assigning singular causes ([318]). We tend to explain particular instances in terms of events that were temporally contiguous (if not continuously linked) so as to determine blame. . Even so, immediacy is only a heuristic, causally linked events can be disentangled in a variety of ways[319].

You can also interpret the role of time and explanation less cognitively. For example, you could see the *Garcia*[160] *effect* as a case of causal attribution. In that case, saccharine – when was appropriately timed with γ-radiation – was a substance was attributed to nausea, resulting in taste aversion[160]. Rats can make some kinds of causal interventional predictions[320], and the dynamics of extinction phenomena can also be modelled using causal structure induction paradigm.[321].

In fact, we can think of much of the premise of the reaction-time experiments as being instances where we(as human scientists needing to make inferences) are attempting to make guesses about features of the world. In particular, we will appeal to the delays that occur not

[§] Or, if we do, it is only after (hypothetically or actually) taking them to be certain. We may explain universal statements, which presuppose continuation into the future, but that is not so different from certainty in the case of general explanation. In that, in certainty in the case of a general explanation, it is claimed to be certain on at least one dimension rather than another. For example, universal in space or time, but localised in particular instances.

just in general as a means of inferring their underlying mental structure, but to the delays that occur for particular individuals so as to infer their particular structure.

If we are right in expressing feedback loops as continuously interacting cycles, then we have little choice than to understand explaining the state of that system in terms of its temporal dynamics. It is difficult to see how you could begin explaining any regulated phenomenon in terms of its dynamics as well as its steady state(s). It is only by virtue of its temporal dynamics that it can be so regulated and distinguished from an unregulated process that happens to take on that state value.

However, note that this points a way out of the apparent paradox of explaining peoples' actions by virtue of their intended outcomes. From a naïve analysis this seems to suggest that we are explaining causes in terms of their effects. However, that simply has not unwrapped the full ontology and event set that actually brought these events about. It is not that we wish to explain a person's actions by their actual outcomes (which follow the action), but by their expected outcomes according to the mental representation of whatever the world was and the causal handles available for bringing about that outcome. To the extent that the person's representation of the the world and its handles are accurate, their ability to execute on their intention is accurate, and the invariability of their intention to means and across time (thus ensuring the validity analysis as a equifinal continuous-time causal feedback loop) it will seem that an analysis in terms of explaining causes by virtue of their effects would be validated. However, that apparent success is an illusion made possible only because of the tight coördination of these various features.

4.2.1 Time in domain specific causal explanations

I illustrated the importance of temporal information in real-life causal explanation through the case-studies book-ending the models and experiments in Chapter 3. However, there are other real world cases where causal explanation occurs where timing information is crucial.

For example, in cases of legal attribution, the state of mind at the time of actions in question are what matters. The triple damages that arise from knowingly violating a patent only apply if violates are "egregious cases typified by willful misconduct" (on the basis of one's knowledge at the time). Learning about the patent in the future (e.g., after the suit has been filed) does not activate the triple penalty. But this future past asymmetry is not where it ends in legal cases — events not only be in the past, but often their recency or even coöcurrence will be required. We blame a person for being a "drunk driver" because of their intoxication at the time that they

were driving; it is not enough that someone ever has imbibed alcohol. On the other hand, in the case of environmental liability and occupational disease, constrained temporal locality may be proof against guilt. In those cases, persistent exposure may be part of the criterion used to determine a workplace's fault in causing the disease[136].

One could apply similar lines of reasoning to other domains where identifying causally responsible agents has been deemed to be of paramount importance. Fault detection, engineering, manufacturing, forensics, mining¶, insurance, and history* involve people explaining the world and (sometimes) intervening on the world in order to produce particular intended outcomes and thereby being responsible for those outcomes. Formally understanding what it means for people to pursue these endeavours will require a model of how people explain events with respect to different sorts of timing information.

4.3 Induction & Time

Though Chapter 3 covers a great deal on the interaction between causal induction and temporal information, there are some issues that we have not at all addressed.

In 1959, Luce[310] laid out one of the central problems that continues to beguile any theory of human cognition (though it is particularly poignant for computational-level, and rational analyses that take a Bayesian form):

> [T]here seems to be have been an implicit assumption that no difficulty is encountered in deciding among what it is that an organism makes its choices. Actually, in practice, it is extremely difficult to know and much experimental technique is devoted to arranging matters so that the organism and the experimenter are (thought to be) in agreement about what the alternatives are. All of our procedures for data collection and analysis require that the experimenter to make explicit decisions about whether an action did or did not occur, and all of our choice theories… begin with the assumption that we have a mathematically well defined set… How these sets come to be defined for organisms, how they may or may not change with experience, how to detect such changes, etc., are questions that have received but little illumination so far.

¶ In that case, the causal agent would be the resource which is presumed to account for whatever surface signal indicates that the area is worthy of investigation.

* This is not an exhaustive list by any means.

Building a hypothesis space is the version of this problem that applies to the Bayesian modeler as well as the cognitive agent. However, as he notes the problem is that the hypothesis space in any real system is not constant. The number of options available is constantly in flux.

In Chapter 3 we elaborate on how we can express theories that take the role of continuous time seriously. To that extent we have theories whose values change depending on the values of events that occur in the world. However, we treat the inference problem as one where there is a static theory that accounts for the observed data that happens to be dynamic. Induction in the case where a theory itself is varying over time or dependent upon the occurrence of various events is an interesting challenge. I believe that in learning how to address that challenge we will discover steps toward addressing the problems discussed in subsection 4.1.3 and minor section 4.1.3.4.

His note on experimental technique as a means of organising a causal scenario is itself intriguing. It suggests the possibility that experiment design itself could be studied as an object of a computational-level analysis. Indeed, especially if we consider the coördinative nature of his account of experiment design it could be useful to have some mechanism for automatically ensuring that the coördination will work in a way that matches the prior beliefs of the average participant (which Suchow et al.[322] have shown to be possible). Such a mechanism would seem especially appropriate for analysing the quality of experiments that can be entirely formally described such as experiments that are run entirely online using standardised platforms (such as Wallace[323]).

4.4 Explanation, Induction & Time: Building the theory engine

Our understanding of the human mind has been bolstered by efforts to rebuild the basis out of which it arises. This requires incorporating advances in machine learning and artificial intelligence. At the core of my work is the hope that, our comprehension of the building blocks of the mind's capabilities can be extricated not by rebuilding the pieces of the mind itself, but by more carefully characterising what it is that the mind so excels at that allows accomplishing so much.

It is clear that one key to this excellence is the mind's facility with causal-theory based inferences. Theories like these allow the mind to postulate the existence of hidden entities and thereby reasoning of them. That variety of reasoning has been invaluable throughout the history of science, technology, and medicine. Furthermore, we can not only reason about these phenomena and induce generalised theories and models, but we can take these general theo-

ries and models and apply them to particular cases to explain phenomena. Aside from feeding back into the inference process itself, explanations like these can guide intervention as well as further information search.

However, as clear as the empirical research on these topics may be, equally important is the precision afforded by computational modelling. It is this that will allow building a theory engine that captures the relevant structure of the mind's operation. Mechanical engines proved useful when they could play the same role as some previously used power source. It did not matter whether the engine performed the event in the exact same manner as the power source that was currently in use[✠]. All that mattered was that the engine could accept inputs of the right form, operate on those inputs to transform the relevant parts of the input to produce an output of the right form. Engines then can be seen as real-world instances of computational-level/rational analysis for physical systems.

In this case the theory engine will be built not out of mechanical parts but mathematical, logical, and computational parts. Defining the mathematical, logical, probabilistic and computational aspects of the work are crucial steps toward defining the overall structure of the problem space that the human mind is solving. But those parts cannot be assured to be of use to a theory engine without assessing their success in solving the same problems we need to analyse our accounts' performance on modelling human cognitive behaviour on a variety of tasks. When the phenomena in question seem to violate standard assumptions about how the problem is to be approached(as is the case for the studies in Chapter 2), it can be more important to characterise human cognitive behaviour appropriately with an eye toward later computational development. When there are multiple available computational models that differ in their predictions (as in Chapter 1) the engine should be able to accommodate each of these parts (as can be observed in this Explanation Engine). If introducing new computational machinery (as in Chapter 3) testing that the machinery accords with a great variety of the supported use cases will be important for validating the appropriateness and relevance of the new machinery.

4.4.1 Alternatives to the mind as theory engine account

But though we describe this as a programme for building a theory engine, it is the mind that will be the inspiration, source, and benchmark against which we would evaluate any proposed design. It is worth noting that in doing so it means that we treat the questions that can be asked of these generative models as questions that can also be asked of the human mind. We

[✠] In fact, to do so would likely have made it far *less* useful an invention.

expect the model to be able to generate and evaluate causal explanations given a fully specified causal graph. We expect it to be able to rate whether something has a particular causal form, where a role falls on a bidirectional scale with two different functional forms (generation and prevention) at opposite ends, and estimates over a simplex over three potential functional forms(generation, prevention, and null). We expect it to be able to infer causal structure among a set of known mutually inclusive or exclusive possibilities as well as to infer hidden causal structure. This needs to occur with data expressed as combinations of linguistic specification and relative frequencies, sequentially presented collections of individuated binary variables, verbally summarised rates, graphically summarised data of many samples of simultaneous one-shot occurrences over sequential time steps, series of independent one-shot trials, or continuous streams of event occurrences.

Data that range across such a variety of media are not typically found together. Indeed, real-time and textually represented data were often presumed to be analysed using different cognitive mechanisms [190,157,138,2]. It is a testament to the power of the CTCT framework that it can both formulate and successfully explain both of these kinds of data. It suggests that at least as far as reasoning about continuous-time causal relationships a theory that proposes a common underlying theory engine like that I have been describing and that does not treat these as isolable problems will be a preferable account to more modular accounts.

Additionally, if we consider the empirical groundings of our work as stemming back to Hume's [37] associationist psychology or Mill's [38] methods of inference, my theory engine theory also performs better. At the least it encompasses a much wider range of phenomena than either of those begin to cover. Of course, I am aided in that I do so with greater mathematical precision and with greater computational power at my disposal. Even advanced versions of these associative theories fail to account for much causal induction phenomena even in the case of binary events occurring over discrete trials (see discussion in Griffiths and Tenenbaum [30], Buehner [138]). And even those theoretical approaches that would be more amenable to our cognitivistic approach (such as the work by Gallistel and Gibbon [155], Gallistel et al. [324]) cannot cover as wide a range of human causal inferential phenomena as I do[⊎].

For the Behaviourist interpretations of this empiricist mantle, at least at a first glance we are not even studying a proper scientific topic. Explanation and induction are mental events that we happen to be eliciting in a standardised form, but because they are mental events from a Behaviourist perspective they are not objects of scientific inquiry. Worse from the Behaviourist, is that the objects about which some of these hidden events and entities are reasoning are

[⊎] They more than make up for this in their coverage of non-human inferential behaviour.

themselves hidden events and entities.

But if the aim is empirical and even mathematical or computational rigour, by all standards my work goes beyond what was available to the Behaviourists[107,106] even if they had been at the cutting-edge of mathematical, probabilistic and computational techniques. I am creating models that involve far more intricately defined state spaces with more variables interacting in far more complicated ways. My predictions are rooted not on the basis of schedules defined in terms of average relative frequency terms or rates. Instead, I have available to me exact occurrences, discrete approximations to those exact occurrences and total counts of relative occurrences. I can incorporate independence in the form of explicit trials or continuous event streams as part of the formal framework itself, rather than needing to presuppose in order to determine what exact kind of "black-box"-style statistical analysis.

The types of responses received from people are directly mappable to the kinds of computational operations and outputs made by our normative models of inference. They may not be *uniquely* mappable (consider subsection 3.13.5), but that does not stop the available maps from being readily interpretable; sometimes more than one proves to be a good model of the phenomenon in question.

If the available models diverge strongly in their predictions, we would not expect all those models to fare well (as can be seen in Chapter 1). But when the models differ strongly in their degree of support, these resulting differences are informative as to the underlying features that make the successful models successful (in the domains in which they are successful). Often, because they are formed out of a rich structure, these differences are informative in ways that the modelling frameworks stemming from other behaviorist and associationist approaches cannot match.

My success stems not merely from using a (purportedly unnecessary[106]) theory but by accounting for people's responses through models that themselves explicitly framed in terms of theories. It is difficult to see how any of what I have accomplished could be done without this explicit reliance on logically constructed theories. They allow a means of representing the hierarchically organised knowledge embedded in (unobservable) human theories. With inferential practices, particular responses and explicitly defined stimuli, this characterises the structure of the total problem people are reasoning about almost wholly in terms of unobservable entities.

4.4.2 Behaviourism and the computational level: the making of a Cognitive Behaviourist

Few cognitive scientists doubt that mental events are objects worthy of study — to do so would almost seem to be a contradiction. Accordingly, Behaviourism is not particularly lauded in the cognitive sciences. But there were valuable features embedded in logical-positivist infected rhetoric: notably a great care in describing the data that they input to their subjects as stimuli and the outputs that they received (even if they did not have particularly careful ways of analysing that data). In a sense, this work – particularly Chapter 3 – is an exercise in taking the best lessons available from the Behaviourist approach while letting the ontological baggage that weighed down their research programme slide through my fingers, never to be seen or heard from again.

If we do not precisely characterising the data available to people in all of its richness, we will be blind to the problems that people are actually solving. Though this is a core principle in computational level and rational analyses, it has been emphasised less than the accurate characterisation of the problem space itself[31,32]. I have benefited greatly from the amount of concern earlier researchers put in appropriately characterising the problem space. But with the Behaviourists' heightened concern for the nature of the input and output data conjoined with the Marrian concern for casting higher-level cognition in the appropriate form, I deliver on the computational-level promise.

In a sense, the computational level does not concern itself with the internal features of the cognitive system that make it cognitive; the computational level cares only about the structure of the problem the system solves. In fact, there need be nothing particularly *cognitive* about the system *other than* the structured nature of the problem it solves. That problem structure would be a bare skeleton without the data to give flesh to its features. From that perspective, a computational-level account will be best understood when it is defined to operate on precisely characterised inputs that it transforms into precisely characterised outputs. That holds regardless of the physical nature of the system in question.

In that case, Marr (in his most computational mode) and Skinner (in his most empiricist mode) – and their intellectual progeny – differ less in terms of methodology and more in terms of their mathematical sophistication. The most extreme version of computational-level analysis is one joined with behaviourist tendencies; it leads to Cognitive Behaviourism.

It is for this reason that I can justify the claim that I am reverse engineering the mind as a theory engine when, in more precise terms, I am writing probabilistic programs inspired by the

conceptual account of a theory engine modelled after the computational structure of human causal theories constrained by the aim to describe human causal cognitive behaviour. In the ideal case, humans would be able to interact with the cognitive behavioural system and communicate with it as facilely as they do one one another. The theory engine would be able to read a textbook and extract from it the abstract knowledge structures that makes the textbook worth publishing. Information would pass between these cognitive behavioural systems, inferences would be provided for one another, explanations and data would be proffered, accepted and assimilated, and new data would be collected in light of all this. The resulting system is ultimately as mutually beneficial as any computational programme aimed at reproducing aspects of higher-order human cognition can be. Such success is the promise of Cognitive Behaviourism.

5

Epilogue: The mind as theory engine

If we wish to build a causal theory engine that even begins to approach the powers of the mind in these matters, we need to address both explanation and induction. If we were able to build such an engine, it could be expected to take causal theories as input to power its basic functioning (to consume theories as a gasoline engine consumes gasoline), to distil causal theories into more useful forms (to process theories as the cotton 'gin removed bolls from lint) and to produce causal theories from raw materials such as data or preprocessed derivatives from other theories (to generate theories as a automotive engine generates motion).

5.1 The human mind is an engine that consumes theories

The mind is capable of learning from the theoretical abstractions gained from others' experience and incorporating them with their own experience. This allows accruing knowledge over time in the form of science and history while also ensuring that the knowledge is available to be used in people's everyday lives. When events occur that fit within the theories one has acquired this will often be more than enough to figure out what one should do with that data. By virtue of the causal theories you have accumulated evidence that can be explained, accounted for and dealt with. Furthermore, by having the ability to rapidly reproduce and apply the carefully considered summaries of others' experience introduces conceptual structures that may allow one's own data to be seen in a new light.

5.2 The human mind is an engine that processes theories

The mind is capable of taking theoretical accounts of the same set of data and determining their relative qualities and abstracting over those qualities across many instances. With a collection of causal theories, one can derive those aspects that are most useful for a domain (in that they are features that many theories in similar domains share). That in turn can be used as an input to new theories for that domain, even if the abstracted piece is itself not a complete theory, but only a theory fragment. A large collection of theories also allows identifying the most distinctive features of theories. Once distinctive features of theories are identified, their role of those features can be learned during the course of using the theories; building a collection of theory modifications that can be appended to existing or new theories if they are needed for similar tasks. This allows transferring knowledge across domains and between individuals that allows improve existing theories with minimal modifications. This will allow memoizing an ontology, making categorization possible both by direct inference and by extraction and application between theories.

5.3 The human mind is an engine that generates theories

The mind is capable of taking a set of experiential data and extracting from that abstract generalities that can support causal action and inference in new situations. Because these theories carry with them extensive structure and strong claims about how it is that the world works, causal theories enable powerful inferences to be made even with small amounts of data. These theories allow people to infer hidden mechanisms in the world that explain observations; this includes new observations that are accounted for by the generalised application of theories shown to be effective in the past. Then, when new anomalies arise that do not fit in with the current theoretical framework, we are emminently capable of developing new conceptual structures and investigating unknown parts of the world until we are satisified with our understanding about just what occurred. The explanations we tell ourselves and others allow this newfound knowledge to be rapidly disseminated and (more usefully) corrected when it comes into contact and conflict with others' similarly generated theories.

A

Simplicity: Additional Materials and Results

A.1 Materials

A.1.1 Full text, Stimuli from Experiments 2 and 4 (Diamond-Structure)

Notes:

Experiment 1 and 3 involved very similar materials. However, Experiment 1 did not include the information about the diagnostic tests, and in Experiment 3 the task order varied across conditions. "Symptom_1" and "Symptom_2" are placeholders for named symptoms.

Page breaks are designated with a triple-horizontal rule like the following:

PLEASE READ THE FOLLOWING INSTRUCTIONS CAREFULLY
ONCE YOU CLICK NEXT YOU CANNOT GO BACK.
Make sure you have read everything carefully before clicking next.
There is a population of aliens that lives on planet Zorg. You are a doctor trying to understand alien medical problems.

Morad's disease and Tritchet's disease together always cause symptom_1 and symptom_2 . If either disease is not present, neither symptom will occur.

One of several ways to contract Tritchet's disease and Morad's disease is to first develop Hummel's disease, which causes both Tritchet's disease and Morad's disease. Hummel's can only cause both of these diseases or neither of them. It will never cause just Morad's disease or just Tritchet's disease.

Aliens can also develop Tritchet's disease and/or Morad's disease independently of having Hummel's disease.

Nothing else is known to cause symptom_1 and symptom_2, i.e. only aliens who have Tritchet's and Morad's disease develop symptom_1 and symptom_2.

Is it possible to develop Tritchet's disease or Morad's disease without having Hummel's disease?

☐ Yes

☐ No

Is it possible to develop symptom_1 and symptom_2 without having Tritchet's disease and Morad's disease?

☐ Yes

☐ No

Unfortunately the particular incidence rates of these diseases are unknown. In order to address this issue, the hospital you work in is running diagnostic tests on a random sample of the population.

The diagnosis machines have 3 lights as shown below, if the H light is yellow, that means the alien has Hummel's disease, if the M light is yellow that means the alien has Morad's disease, and if the T box is yellow that means the alien has Tritchet's disease.

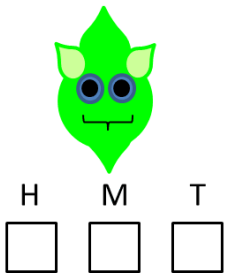
If a box is empty, that means the alien does not have the respective disease. The lights only turn on (turn yellow) if the alien has the disease.

The hospital only has a limited number of diagnosis machines, so they have to test aliens in a series of groups to get the full sample. Each group of aliens will step into the diagnosis machines, which will read out whether they have each disease.

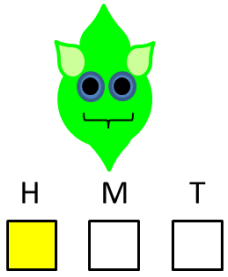
Each group of aliens will stay in the machines for a few seconds. Then those aliens will exit the machines as a new group of aliens enters the machines, which also takes a few seconds. Then the machines will turn on and you will get information about the new group of aliens. This will happen several times in order for you to see the full sample. This may appear as if your screen is refreshing -- each time this happens it is the result of a new group of aliens either

leaving or entering the diagnostics machines.

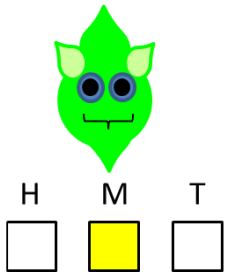
Below are the various outcomes that the diagnosis machine could produce. Note some of these outcomes may not be present in the actual data because they are impossible given the way the diseases work.



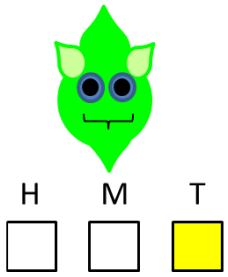
This alien would have no diseases.



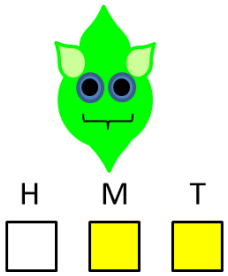
This alien would have only Hummel's disease.



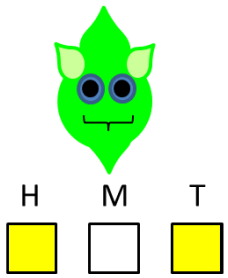
This alien would have only Morad's disease.



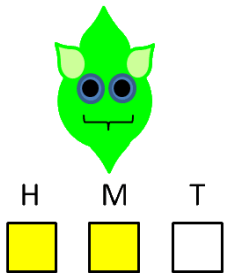
This alien has only Tritchet's disease.



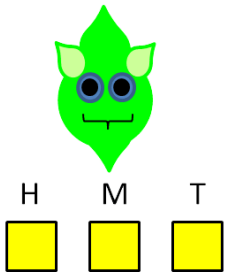
This alien would have Morad's and Tritchet's disease, but not Hummel's disease.



This alien would have Hummel's and Tritchet's disease, but not Morad's disease.

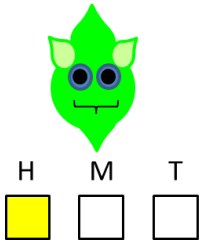


This alien would have Hummel's and Morad's disease, but not Tritchet's disease.



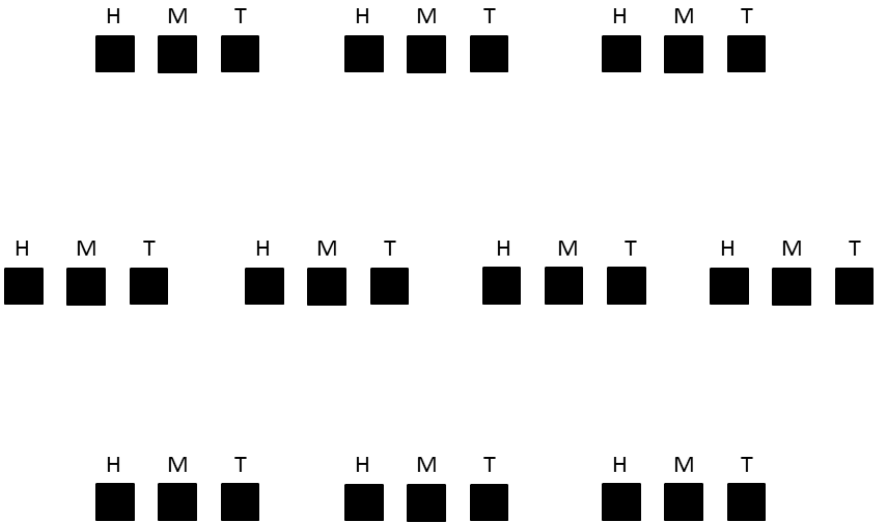
This alien would have all three diseases.

Just to demonstrate that you understand how the machines work, if you saw the following read-out when an alien is in a machine that would mean the alien had which disease(s)?



☐ Hummel's disease and Morad's Disease, but not Tritchet's disease
☐ Morad's disease and Tritchet's disease but not Hummel's disease
☐ No diseases
☐ Only Hummel's disease
☐ Only Morad's disease
☐ Only Tritchet's disease
☐ Hummel's disease and Tritchet's disease, but not Morad's disease
☐ All three diseases

Here is the room of diagnostic machines that you will be working with. There are currently no aliens in the machines, this is only to give you an idea of how the room is laid out.



Now that you have seen the empty room, we'll let the aliens into the machines. You will see the results of the diagnostic tests on a random sample from the population. Remember, you will see the results from different groups of aliens presented one after the other.

Together, all of the groups make up the entire random sample, and **no alien appears in more than one group**.

Each group will only appear for a few seconds.

Remember: when the screen refreshes, that is just a new group of aliens moving into the machine and the results of this new groups' diagnoses.
Now an alien, Treda, comes to you. Treda has two symptoms: Treda has symptom_1 and symptom_2
What do you think is the **most satisfying explanation** for the symptoms Treda is exhibiting?

☐ Treda **has Hummel's disease**, which **caused Tritchet's disease and Morad's disease**, which together caused symptom_1 and symptom_2 .

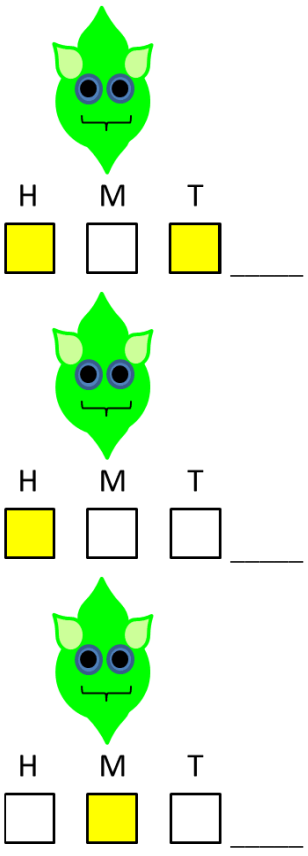
☐ Treda developed symptom_1 and symptom_2 , but has **none** of the aforementioned diseases.

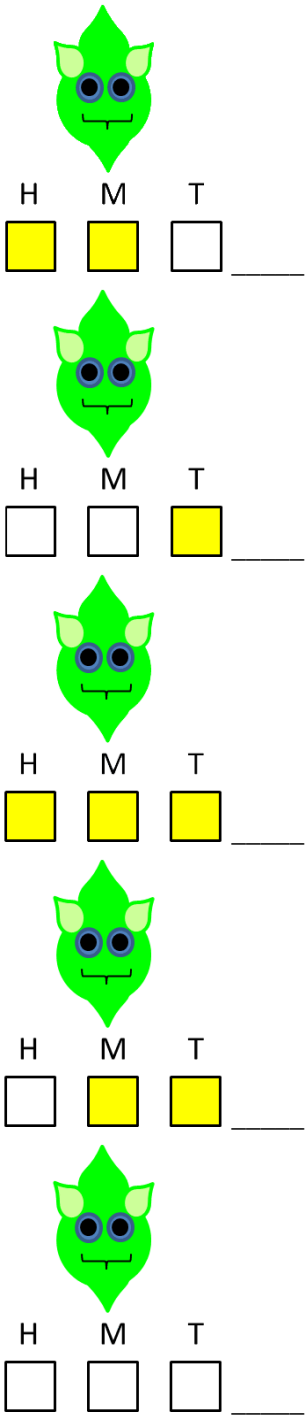
☐ Treda **does not have Hummel's disease**, and **independently developed Tritchet's disease and Morad's disease**, which together caused the symptom_1 and symptom_2 .

Why did you choose this explanation?

Think back to the series of aliens that you saw in the diagnostics machines. There were 120 aliens total. How many aliens of the 120 had diagnoses that correspond to the following images?

Keep in mind that the numbers should total 120.





Do you think Treda had…

	Yes	No
Tritchet's Disease	☐	☐
Hummel's Disease	☐	☐
Morad's Disease	☐	☐

Does Hummel's disease cause Tritchet's{and Morad's} disease?

☐ **Yes**

☐ **No**

Does Hummel's disease prevent Tritchet's{and Morad's} disease?

☐ **Yes**

☐ **No**

What is your age?____

What is your sex?

☐ Male

☐ Female

☐ Other / Prefer not to specify

Most modern theories in psychology recognize the fact that people do not think in a vacuum. Individual preferences and knowledge, along with situational variables, can greatly impact their behavior. In order to facilitate the research, we are interested in knowing certain factors about you, the decision maker. Specifically, we are interested in how closely you read the directions; if not, then some of our experiments that rely on changes in the instructions will be invalid. So, to demonstrate that you have read the instructions, please ignore the sports items below and write in the box that says other that you have read the instructions. Answering the material in this question provides more information, and having more data about participants is helpful for conducting the experiment. Thank you for entering the requested information.

- Skiing
- Soccer
- Running
- Baseball
- Football
- Tennis
- Basketball
- Swimming
- Other

Thank you for taking this study. Before you finish, if you have any comments regarding this study (e.g. if you saw any gigantic errors or inconsistencies, whether you had any technical problems viewing it, or anything else that comes to mind) please include them in the blank below. Click the arrow when you are finished in order to complete the study.

A.1.2 Chain structure modifications.

Any cases in which Morad's appears, eliminate reference to Morad's.

E.g.,

This:

Morad's disease and Tritchet's disease together always cause symptom_1 and symptom_2 . If either disease is not present, neither symptom will occur.

Becomes:

Tritchet's disease always causes symptom_1 and symptom_2 . If the disease is not present, neither symptom will occur.

And this:

Do you think Treda had…

	Yes	No
Tritchet's Disease	☐	☐
Hummel's Disease	☐	☐
Morad's Disease	☐	☐

Becomes:

Do you think Treda had…

	Yes	No
Tritchet's Disease	☐	☐
Hummel's Disease	☐	☐

And so on…

A.2 Reading/Comprehension Checks

[1,2,3,4] ← designates which experiments these were used in

[*]← means used in all 4 experiments, as is appropriate given other conditions.

⊙← indicates the correct answer choice.

A.2.1 Exclusion criteria

A.2.1.1 Necessity-Check [*]

Is it possible to develop symptom_1 and symptom_2 without having Tritchet's disease { *and* Morad's disease}?

☐ **Yes**

⊙ **No**

A.2.1.2 Non-necessity-Check [*]

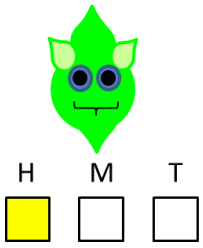
Is it possible to develop Tritchet's disease {or Morad's disease} without having Hummel's disease?

⊙ **Yes**

☐ **No**

A.2.1.3 Machine-Comprehension-Check [2–4]

Just to demonstrate that you understand how the machines work, if you saw the following read-out when an alien is in a machine that would mean the alien had which disease(s)?



☐ Hummel's disease and Morad's Disease, but not Tritchet's disease

☐ Morad's disease and Tritchet's disease but not Hummel's disease

☐ No diseases

⊙ Only Hummel's disease

☐ Only Morad's disease
☐ Only Tritchet's disease
☐ Hummel's disease and Tritchet's disease, but not Morad's disease
☐ All three diseases

A.2.1.4 Disease-Count-Check[2–4]:

Did their estimated counts of the diseases sum to 120 (as instructed) or 100 (under a probabilistic interpretation of the question).

A.2.1.5 Explanation-Choice-Other[*]:

> *Note: In the rare cases where a participant chose "Treda developed symptom_1 and symptom_2 , but has none of the aforementioned diseases.", because that directly conflicted with the information given earlier, we took this answer to mean that they misunderstood the scenario and thus were excluded from analysis.*

Now an alien, Treda, comes to you. Treda has two symptoms: Treda has symptom_1 and symptom_2

What do you think is the **most satisfying explanation** for the symptoms Treda is exhibiting?

☐ Treda **has Hummel's disease**, which **caused Tritchet's disease and Morad's disease**, which together caused the symptom_1 and symptom_2 .

☐ Treda developed symptom_1 and symptom_2 , but has **none** of the aforementioned diseases.

☐ Treda **does not have Hummel's disease**, and **independently developed Tritchet's disease and Morad's disease**, which together caused the symptom_1 and symptom_2 .

A.2.1.6 Inference-Check: [*]

> *Note: This question was included to ensure that participants understood that the presence of Treda's two symptoms implied the presence of Tritchet's and Morad's diseases; no exclusions were made on the basis of response to the question about Hummel's.*

Do you think Treda had…

	Yes	No
Tritchet's Disease	⊙	☐
Hummel's Disease	?	?
Morad's Disease	⊙	☐

A.2.1.7 Cause-Check: [*]

Does Hummel's disease cause Tritchet's{and Morad's} disease?

⊙ **Yes**

☐ **No**

A.2.1.8 Prevent-Check: [*]

Does Hummel's disease prevent Tritchet's{and Morad's} disease?

☐ **Yes**

⊙ **No**

A.2.1.9 Instructional-Manipulation-Check (IMC): [*]

Most modern theories in psychology recognize the fact that people do not think in a vacuum. Individual preferences and knowledge, along with situational variables, can greatly impact their behavior. In order to facilitate the research, we are interested in knowing certain factors about you, the decision maker. Specifically, we are interested in how closely you read the directions; if not, then some of our experiments that rely on changes in the instructions will be invalid. So, to demonstrate that you have read the instructions, please ignore the sports items below and write in the box that says other that you have read the instructions. Answering the material in this question provides more information, and having more data about participants is helpful for conducting the experiment. Thank you for entering the requested information.

- Skiing
- Soccer
- Running
- Baseball
- Football
- Tennis
- Basketball
- Swimming
- Other

A.2.2 Misunderstood Explanation Justification Criterion

Explanation justifications that suggested the participant misunderstood some aspect of the experiment were classified as "misunderstood," and participants whose explanations fell into this category were excluded from additional analyses.

A.2.3 Proportions of each exclusion criterion split by experiment

Experiment	Neccessity	Non-necessity	Diagnosis Machine
Exp 1	0.191176471	0.176470588	N/A
Exp 2	0.205217391	0.099130435	0.022608696
Exp 3	0.264781491	0.100257069	0.017994859
Exp 4	0.195121951	0.082926829	0.029268293

Experiment	Cause Check	Prevent Check	Disease count check
Exp 1	0.147058824	0.044117647	N/A
Exp 2	0.189565217	0.024347826	0.050434783
Exp 3	0.231362468	0.035989717	0.051413882
Exp 4	0.126829268	0.019512195	0.063414634

Experiment	Inference Check	IMC	Misunderstood
Exp 1	0.132352941	0.029411765	0.073529412
Exp 2	0.088695652	0.062608696	0.097391304
Exp 3	0.102827763	0	0.138817481
Exp 4	0.058536585	0	0.092682927

A.3 Explanation Justifications Experiments 2–4

A.3.1 Experiment 2.

Explanation choice justifications were coded as in Experiment 1. There was moderate agreement amongst the raters (returning all instances of "Misunderstood" to the dataset that were

not excluded for other reasons; Fleiss $\kappa = 0.4415, z = 29.46, p < 10^{-4}$. The distribution of explanation justifications can be found in Table A.1. We found a significant difference between the overall justification distributions across Causal Structures, $\chi^2(308) = 8.7738, p < .05$, with participants more likely to invoke probability in Chain-Structure than in Diamond-Structure.

As in Experiment 1, the proportion of justifications that appealed to simplicity was quite small (8%, N = 25). Of these, fourteen were used to support the proximal-choice in the Chain-Structure condition, zero to support the complete-choice in the Chain-Structure condition, eight to support the proximal-choice in the Diamond-Structure, condition, and three to support the complete-choice in the Diamond-Structure condition.

Table A.1: Proportions of justification types by condition, Exp 2.

	Overall	Chain-Structure	Diamond-Structure
Simplicity:	8.0%	8.9%	7.2%
Probability:	52.4%	58.2%	46%
Other:	33.1%	29.8%	36.6%
Misunderstood:	6.4%	3.1%	9.8%

A.3.2 Explanation Choice Justifications in Experiment 3.

Justifications were coded as in Experiments 1–2, yielding moderate agreement among coders ($\kappa = .5768, z = 35.58, p < 10^{-4}$). The justifications distributions differed between the explain-first and the estimate-first conditions, $\chi^2(185) = 7.9078, p < 0.05$, with participants more likely to provide Other justifications in *estimate-first* (see Table A.2). As in Experiments 1–2, the proportion of justifications that appealed to simplicity was quite small ($4.8\%, N = 9$), with the following distribution across conditions and explanation choices: two were used to support the proximal-choice in the *explain-first* condition, two to support the complete-choice in the *explain-first* condition, three to support the proximal-choice in the *estimate-first* condition, and two to support the complete-choice in the *estimate-first* condition.

A.3.3 Explanation Choice Justifications in Experiment 4.

Justifications were coded as in Experiments 1–3, with substantial agreement between the three raters ($\kappa = 0.7484, z = 24.538, p = 10^{-4}$). Overall, justifications invoked simplicity in

Table A.2: Proportion of justification types by condition, Exp 3.

	Overall	*Explain-first*	*Estimate-first*
Simplicity:	4.8%	4.2%	5.4%
Probability:	43.6%	49.0%	38.0%
Other:	40.9%	32.3%	50%
Misunderstood:	10.6%	14.6%	6.5%

1.6% of cases, probability in 52.9%, and other justifications in 40.7%. The remaining 4.9% of participants who passed other reading checks provided explanations that were designated as misunderstood, and were therefore excluded from other analyses. There were two people who justified their explanation choice with reference to simplicity; one who chose complete, one who chose proximal.

B

An introduction to Causal Bayesian Networkx(cbnx)

B.1 Introduction and Aims

My first goal in this appendix is to provide enough of an introduction to some formal and mathematical tools such that those familiar with `python` and programming more generally will be able to appreciate both why and how one might implement causal Bayesian networks. Especially to exhibit *how*, I have developed parts of a toolkit that allows the creation of these models on top of the `NetworkX` python package. Given the coincidence of the names, it seemed most apt to refer to this toolkit as `Causal Bayesian NetworkX` abbreviated as cbnx*

If you wish to see an application of this work to the problem discussed in section 3.12, you can view the repository of the relevant code on GitHub at hidden_structure_inference. This code currently builds a forward sampling based modeling system that avoids using the *detective probability model* approximation discussed in subsection 3.12.5. However, it is only capable of handling the case with zero base-rates, which means that it is unable to capture human data well. As discussed in minor section 3.12.6.2 non-zero baserates are needed to have a

* Static code can be found in the document from which the content of this appendix is drawn Pacer [325], and the most recent version of the code can be found at cbnx. cbnx is licensed with the BSD 3-clause license.

smooth distribution over graphs; otherwise, like in the detective model, the forward sampling approach will find that almost all samples for almost all graphs will be given 0 likelihood (or $-\infty$ loglikelihood).

In order to understand the tool-set requires the basics of probabilistic graphical models, which requires understanding some graph theory and some probability theory. The first few pages are devoted to providing necessary background and illustrative cases for conveying that understanding.

Notably, contrary to how Bayesian networks are commonly introduced, I say relatively little about inference from observed data. This is intentional, as is this discussion of it. Many of the most trenchant problems with Bayesian networks are found in critiques of their use to infer these networks from observed data. But, many of the aspects of Bayesian networks (especially causal Bayesian networks) that are most useful for thinking about problems of structure and probabilistic relations do not rely on inference from observed data. In fact, I think the immediate focus on inference has greatly hampered widespread understanding of the power and representative capacity of this class of models. Equally – if not more – importantly, I aim to discuss generalizations of Bayesian networks such as those that appear in Griffiths and Tenenbaum [1], and inference in these cases requires a much longer treatment (if a comprehensive treatment can be provided at all). If you are dissatisfied with this approach and wish to read a more conventional introduction to (causal) Bayesian networks I suggest consulting Pearl [28].

The current instantiation of the CBNX toolkit can be seen as consisting of two main parts: graph enumeration/filtering and the storage and use of probabilistic graphical models in a NetworkX compatible format [4].

I focus first on establishing a means of building iterators over sets of directed graphs. I then apply operations to those sets. Beginning with the complete directed graph, we enumerate over the subgraphs of that complete graph and enforce graph theoretic conditions such as acyclicity over the entire graph, guarantees on paths between nodes that are known to be able to communicate with one another, or orphan-hood for individual nodes known to have no parents. We accomplish this by using closures that take graphs as their input along with any explicitly defined arguments needed to define the exact desired conditions.

I then shift focus to a case where there is a specific known directed acyclic graph that is imbued with a simple probabilistic semantics over its nodes and edges, also known as a Bayesian network. I demonstrate how to sample independent trials from these variables in a way consistent with these semantics. I discuss some of the challenges of encoding these semantics in dictionaries as afforded by NetworkX without resorting to `eval` statements.

I conclude by discussing Computational Cognitive Science as it relates to graphical models and machine learning in general. In particular, I will discuss a framework called **theory based causal induction**[1], or my preferred term: **causal theories**, which allows for defining problems of causal induction. The perspective expressed in this appendix, the associated talk, and the CBNX toolkit developed out of this framework.

B.1.1 Graphical Models

Graphs are defined by a set of nodes (X, $|X| = N$) and a set of edges between those nodes ($E|e \in E \equiv e \in (X \times X)$).

B.1.1.1 Notes on notation

Nodes In the examples in CBNX, nodes are given explicit labels individuating them such as $\{A, B, C, ...\}$ or {'rain', 'sprinkler', 'ground'}. Often, for the purposes of mathematical notation, it is better to index nodes with integers over a common variable label, e.g., using $\{X_1, X_2, ...\}$.†

Edges Defined in this way, edges are all *directed* in the sense that an edge from X_1 to X_2 is not the same as the edge from X_2 to X_1, or $(X_1, X_2) \neq (X_2, X_1)$. An edge (X_1, X_2) will sometimes be written as $X_1 \rightarrow X_2$, and the relation may be described using language like "X_1 is the parent of X_2" or "X_2 is the child of X_1".

Directed paths Paths are a useful way to understand sequences of edges and the structure of a graph. Informally, to say there is a path between X_i and X_j is to say that one can start at X_i and by traveling from parent to child along the edges leading out from the node that you are currently at, you can eventually reach X_j.

† Despite pythonic counting beginning with 0, I chose not to begin this series with 0 because when dealing with variables that might be used in statistical regressions, the 0 subscript will have a specific meaning that separates it from the rest of the notation. For example when expressing multivariate regression as $Y = \beta X + \epsilon, \epsilon \sim \mathcal{N}(0, \Sigma)$, β_0 refers to the parameter associated with a constant variable $x_0 = 1$ and X is normally defined as $x_1, x_2, x_3, ...$ This allows a simple additive constant to be estimated, which often is not of interest to statistical tests, acting as a scaling constant. This makes for a simpler notation than $Y = \beta_0 + \beta X + \epsilon$, because that is equivalent to $Y = \beta X + \epsilon$ if $x_0 = 1$. But, in other cases (e.g., Pacer and Griffiths[156]) 0 index will be used to indicate background sources for events in a system.

To define it recursively and more precisely, if the edge (X_i, X_j) is in the edge set or if the edges (X_i, X_k) and (X_k, X_j) are in the edge set there is a path from X_i to X_j. Otherwise, a graph has a path from node X_i to X_j if there is a subset of its set of edges such that the set contains edges (X_i, X_k) and (X_l, X_j) and there is a path from X_k to X_l.

B.1.1.2 Adjacency Matrix Perspective

For a fixed set of nodes X of size N, each graph is uniquely defined by its edge set, which can be seen as a binary $N \times N$ matrix, where each index (i, j) in the matrix is 1 if the graph contains an edge from $X_i \rightarrow X_j$, and 0 if it does not contain such an edge. We will refer to this matrix as $A(G)$.

This means that any values of 1 found on the diagonal of the adjacency matrix (i.e., where $X_i \rightarrow X_j, i = j$) indicate a self-loop on the respective node.

B.1.1.3 Undirected Graphs

We can still have a coherent view of *undirected* graphs, despite the fact that our primitive notion of an edge is that of a *directed* edge. If a graph is undirected, then if it has an edge from $X_i \rightarrow X_j$ then it has an edge from $X_j \rightarrow X_i$. Equivalently, this means that the adjacency matrix of the graph is symmetric, or $A(G) = A(G)^\top$. However from the viewpoint of the undirected graph, that means that it has only a single edge.

B.1.1.4 Directed Graphs

From the adjacency matrix perspective we've been considering, all graphs are technically directed, and undirected graphs are a special case where one (undirected) edge would be represented as two symmetric edges.

The number of directed graphs that can be obtained from a set of nodes of size n can be defined explicitly using the fact that they can be encoded as a unique $n \times n$ matrix:

$$R_n = 2^{n^2}$$

Directed Acyclic Graphs A cycle in a directed graph can be understood as the existence of a path from a node to itself. This can be as simple as a self-loop (i.e., if there is an edge (X_i, X_i) for any node X_i).

Directed acyclic graphs(DAGs) are directed graphs that contain no cycles.

The number of DAGs that obtainable from a set of n noddes can be defined recursively as follows[326]:

$$R_n = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k(n-k)} R_{n-k}$$

Note, because DAGs do not allow any cycles, this means that there can be no self-loops. As a result, every value on the diagonal of a DAG's adjacency matrix will be 0.

B.2 Probability Distributions: Conditional, Joint and Marginal

A random variable defined by a conditional probability distribution[‡] has a distribution indexed by the realization of some other variable (which itself is often a random variable, especially in the context of Bayesian networks).

The probability mass function (pmf) for discrete random variable X with value x will be noted as $P(X = x)$. Often, when discussing the full set of potential values (and not just a single value), we leave out the $= x$ and just indicate $P(X)$.[§]

The conditional probability of X with value x given another variable Y with value y is $P(X = x \,|Y = y)$. Much like above, if we want to consider the probability of each possible event without specifying one, sometimes this will be written as $P(X|Y = y)$. If we are considering conditioning on any of the possible values of the known variable, we might use the notation $P(X|Y)$, but that is a slight abuse of the notation.

You *can* view $P(X|Y)$ as a function over the $X \times Y$ space. But do not interpret that as a probability function. Rather, this defines a probability function for X relative to each value of Y. Without conditioning on Y we have many potential probability functions for X. Equivalently, it denotes a *family* of probability functions on X indexed by the values $Y = y$.

[‡] Rather than choose a particular interpretation of probability over event sets (e.g., Bayesian or frequentist), I will attempt to remain neutral, as those concerns are not central to the issues of graphs and simple sampling.

[§] If one is dealing with continuous quantities rather than discrete quantities one will have to use a probability density function (pdf) which does not have as straightforward an interpretation as a probability mass function. This difficult stems from the fact that (under most cases) the probability of any particular event occurring is "measure zero", or "almost surely" impossible. Without getting into measure theory and the foundation of calculus and continuity we can simply note that it is not that any individual event has non-zero probability, but that sets of events have non-zero probability.As a result, continuous random variables are more easily understood in terms a cumulative density function (cdf), which states not how likely any individual event is, but how likely it is that the event in question is less than a value x. The notation usually given for a cdf of this sort is $F(X \leq x) = \int_{-\infty}^{x} f(u)du$, where $f(u)$ is the associated probability density function.

The *joint probability* of X and Y is the probability that both X and Y occur in the event set in question. This is noted as $P(X, Y)$ or $P(X \cap Y)$ (using the set theoretic intersection operation). Similar to $P(X|Y)$, you *can* view $P(X, Y)$ as a function over the space defined by $X \times Y$. However, $P(X, Y)$ is a probability function in the sense that the sum of $P(X = x, Y = y)$ over all the possible events in the space defined by $(x, y) \in X \times Y$ equals 1.

The *marginal probability* of X is just $P(X)$. The term "marginalization" refers to the notion of summing over values of Y in their joint probability. When probabilities were recorded in probability tables, the sum would be recorded in the *margins*. Formally, this can be stated as $P(X) = \sum_{y \in Y} P(X, Y)$.

B.2.1 Relating conditional and joint probabilities

Conditional probabilities are related to joint probabilities using the following form:

$$P(X|Y = y) = \frac{P(X, Y = y)}{P(Y = y)} = \frac{P(X, Y = y)}{\sum_{x \in X} P(X = x, Y = y)}$$

Equivalently:

$$P(X, Y = y) = P(X|Y = y)P(X)$$

B.2.2 Bayes' Theorem

Bayes' Theorem can be seen as a result of how to relate conditional and joint probabilities. Or more importantly, how to compute the probability of a variable once you know something about some other variable.

Namely, if we want to know $P(X|Y)$ we can transform it into $\frac{P(X,Y)}{\sum_{x \in X} P(X=x,Y)}$, but then can also transform joint probabilities ($P(X, Y)$) into statements about conditional and marginal probabilities ($P(X|Y)P(X)$). This leaves us with

$$P(X|Y) = \frac{P(Y|X)P(X)}{\sum_{x \in X} P(Y|X = x)P(X = x)}$$

B.2.3 Probabilistic Independence

To say that two variables are independent of each other means that knowing/conditioning on the realization of one variable is irrelevant to the distribution of the other variable. This is

equivalent to saying that the joint probability is equal to the multiplication of the probabilities of the two events.

If two variables are conditionally independent, that means that conditional on some set of variables, condition

B.2.4 Example: Marginal Independence ≠ Conditional Independence

Consider the following example:

$$\begin{aligned} X &\sim \text{Bernoulli}_{\{0,1\}}(.5),\ Y \sim \text{Bernoulli}_{\{0,1\}}(.5) \\ Z &= X \oplus Y, \oplus \equiv \text{XOR} \end{aligned}$$

Note that, $X \perp\!\!\!\perp Y$ but $X \not\perp\!\!\!\perp Y | Z$.

B.3 Bayesian Networks

Bayesian networks are a class of graphical models that have particular probabilistic semantics attached to their nodes and edges. This makes them probabilistic graphical models.

In Bayesian networks when a variable is conditioned on the total set of its parents and children, it is conditionally independent of any other variables in the graph. This is known as the "Markov blanket" of that node.[¶]

B.3.1 Common assumptions in Bayesian networks

While there are extensions to these models, a number of assumptions commonly hold.

B.3.1.1 Fixed node set

The network is considered to be comprehensive in the sense that there is a fixed set of n known nodes. This rules out the possibility of hidden/latent variables as being part of the network. From this perspective inducing hidden nodes requires postulating a new graph that is potentially unrelated to the previous graph.

[¶] The word "Markov" refers to Andrei Markov and appears as a prefix to many other terms. It most often indicates that some kind of independence property holds. For example, a Markov chain is a sequence (chain) of variables in which each variable depends only on the value of the immediately preceding and postceding variables in the chain. Properties like this make computation easier.

B.3.1.2 Trial-based events, complete activation and DAG-hood

Within a trial, all events are presumed to occur simultaneously.There is no notion of temporal asynchrony, where one node/variable takes on a value before its children take on a value (even if in reality – i.e., outside the model – that variable is known to occur before its child). Additionally, the probabilistic semantics will be defined over the entirety of the graph which means that one cannot sample a proper subset of the nodes of a graph without marginalizing out and incorporating information from the ignored nodes into the subset in question.

This property also explains why Bayesian networks need to be acyclic. Most of the time when we consider causal cycles in the world the cycle relies on a temporal delay between the causes and their effects to take place. If the cause and its effect is simultaneous, it becomes difficult (if not nonsensical) to determine which is the cause and which is the effect — they seem instead to be mutually definitional. But, as noted above, when sampling in Bayesian networks simultaneity is presumed for *all* of the nodes.

B.3.2 Independence in Bayes Nets

One of the standard ways of describing the relation between the semantics (probability values) and syntax (graphical structure) of Bayesian networks is how graph encodes particular conditional independence assumptions between the nodes of the graph. Indeed, in some cases Bayesian networks merely play the role of a convenient representation for conditional and marginal independence relationships between different variables.

It is the perspective of the graphs as *merely* representing the independence relationships and the focus on inference that leads to the focus on equivalence classes of Bayes nets. The set of graphs $\{A \rightarrow B \rightarrow C,\ A \leftarrow B \rightarrow C,$ and $A \leftarrow B \leftarrow C\}$ represent the same conditional independence relationships, and thus cannot be distinguished on the basis of observational evidence alone. This also leads to the emphasis on finding V-structures or common-cause structures where (at least) two arrows are directed into the same child with no direct link between those parents (e.g., $A \rightarrow B \leftarrow C$). V-structures are observationally distinguishable because any reversing the direction of any of the arrows will alter the conditional independence relations that are guaranteed by the graphical structure.*

* A more thorough analysis of this relation between graph structures and implied conditional independence relations invokes the discussion of *d-separation*. However, d-separation (despite claims that "[t]he intuition behind [it] is simple") is a more subtle concept than it at first appears as it involves both which nodes are observed and the underlying structure.

Though accurate, this eschews important aspects of the semantics distinguishing arrows with different directions when you consider the kinds of values variables take on.

B.3.2.1 Directional semantics between different types of nodes

The conditional distributions of child nodes are usually defined with parameter functions that take as arguments their parents' realizations for that trial. Bayes nets often are used to exclusively represent discrete (usually, binary) nodes the distribution is usually defined as an arbitrary probability distribution associated with the label of it's parent's realization.

If we allow (for example) positive continuous valued nodes to exist in relation to discrete nodes the kind of distributions available to describe relations between these nodes changes depending upon the direction of the arrow. A continuous node taking on positive real values mapping to an arbitrarily labeled binary node taking on values $\{a, b\}$ will require a function that maps from $\mathbb{R} \rightarrow [0, 1]$, where it maps to the probability that the child node takes on (for instance) the value a[✠].However, if the relationship goes the other direction, one would need to have a function that maps from $\{a, b\} \rightarrow \mathbb{R}$. For example, this might be a Gaussian distributions for *a* and *b* ($(\mu_a, \sigma_a), (\mu_b, \sigma_b)$). Regardless of the particular distributions, the key is that the functional form of the distributions are radically different.

B.3.3 Sampling and semantics in Bayes Nets

The procedure we will use to sample from Bayesian networks uses an *active sample set*. This is the set of nodes for which we have well-defined distributions at the time of sampling.

There will always be at least one node in a Bayesian network that has no parents. We will call these nodes *orphans*. To sample a trial from the Bayesian network we begin with the orphans. Because orphans have no parents – in order for the Bayes net to be well-defined – each orphan will have a well-defined probability distribution available for direct sampling. The set of orphans is the first active sample set.

After sampling from all of the orphans, we will take the union of the sets of children of the orphans, and at least one of these nodes will have values sampled for all of its parents. We take the set of orphans whose entire parent-set has sampled values, and sample from the conditional distributions defined relative to their parents' sampled values and make this the *active sample set*.

[✠] If the function maps directly to one of the labeled binary values this can be represented as having probability 1 of mapping to either a or b.

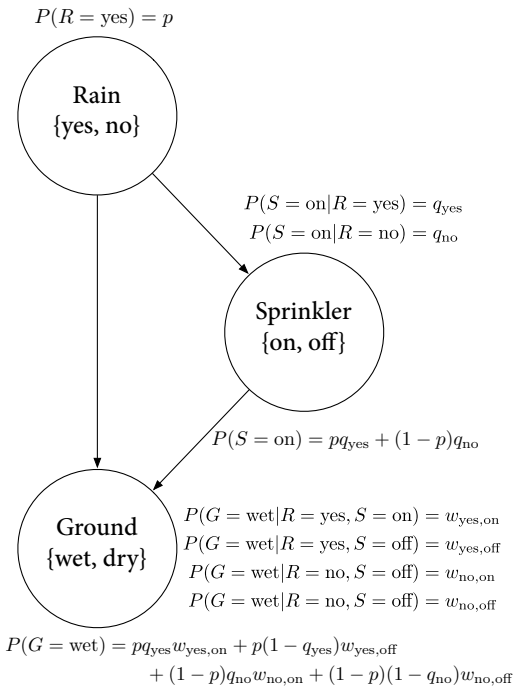


Figure B.1: A Bayesian network describing the sprinkler example. Including both conditional and marginal distributions.

After sampling the active sample set, we will either have new variables whose distributions are well-defined or will have sampled all of the variables in the graph for that trial.

B.3.4 Example: Rain, Sprinkler & Ground

In the sprinkler Bayesian network in Figure B.1[⊎], there three discrete nodes that represent whether it *Rains* (yes or no), whether the *Sprinkler* is on (on or off) and whether the *Ground* is wet (wet or dry). The edges encode the fact that the rain listens to no one, that the rain can alter the probability of whether the sprinkler is on, and the rain and the sprinkler together determine how likely it is that the ground is wet.

[⊎] This is an ill-specified Bayesian network, because while I have specified the states and their relations, I left open the potential interpretation of the parameters and how they relate to one another. I did so because it shows both the limits and strengths of what is encoded knowing only the structure, computing both conditional and marginal distributions for all variables.

B.4 Causal Bayesian Networks

Causal Bayesian networks are Bayesian networks that are given an interventional operation allowing for "graph surgery" by cutting nodes off from their parents[⋉]. Interventions are cases where a causal force is able to exogenously set the values of individual nodes, rendering intervened on nodes independent of their parents.

B.5 NetworkX[4]

NetworkX is a package for using and analyzing graphs and complex networks. It stores different kinds of graphs as variations on a "dict of dicts of dicts" structure. For example, directed graphs are stored as two dict-of-dicts-of-dicts structures[⊛].

B.5.1 Basic NetworkX operations

NetworkX is usually imported using the `nx` abbreviation, and you input nodes and edges as lists of tuples, which can be assigned dictionaries as their last argument, which stores the dictionary as the nodes' or edges' data.

B.6 cbnx: Graphs

Here we will look at some of the basic operations described in the Jupyter notebook found at cbnx. For space and formatting reasons this code may differ slightly from that either in the variable names or comments, for the original version of these code snippets see graph-builder-code.

[⋉] This is technically a more general definition than that given in Pearl [28] as in that case there is a specific semantic flavor given to interventions as they affect the probabilistic semantics of the variables within the network. This is related to his notion of a `do`-operator which deterministically sets a node to a particular value. Because here we are considering a version of intervention that affects the *structure* of a set of graphs rather than an intervention's results on a specific parameterized graph, this greater specificity is unnecessary.

[⊛] It can also represent multi-graphs (graphs where multiple versions of "the same" edge from the adjacency matrix perspective can exist and will (usually) carry different semantics). We will not be using the multigraph feature of NetworkX, as multigraphs are not traditionally used in the context of Bayesian networks.

B.6.0.1 Other packages

In addition to networkX, we need to import numpy[327], scipy[328], and functions from itertools.

```
import numpy as np
import scipy
from itertools import chain, combinations, tee
```

B.6.1 Beginning with a max-graph

Starting with the max graph for a set of nodes (i.e., the graph with N^2 edges), we build an iterator that returns graphs by successively removing subsets of edges. Because we start with the max graph, this procedure will visit all possible subgraphs. One challenge that arises when visiting *all* possible subgraphs is the sheer magnitude of that search space (2^{N^2}).

```
def completeDiGraph(nodes):
    G = nx.DiGraph()
    G.add_nodes_from(nodes)
    edgelist = list(combinations(nodes,2))
    edgelist.extend([(y,x) for x,y in edgelist)
    edgelist.extend([(x,x) for x in nodes])
    G.add_edges_from(edgelist)
    return G
```

B.6.2 Preëmptive Filters

The graph explosion problem is helped by determining which individual edges are known to always be present and which ones are known to never be present. In this way we can reduce the size of the edgeset over which we will be iterating.

Filters can be applied by using `filter_Graph()`, which takes a graph and a filter_set as its arguments and returns a graph. A filter_set is a set of functions that take each take (at least) a graph as an argument and return a graph with a reduced edgeset according to the semantics of the filter.

```
def filter_Graph(G,filter_set):
    graph = G.copy()
    for f in filter_set:
        graph = f(graph)
    return graph
```

B.6.3 Example filter: remove self-loops

By default the graph completed by `completeDiGraph()` will have self-loops, often we will not want this (e.g., DAGs cannot contain self-loops).

```
def extract_remove_self_loops_filter():
    def remove_self_loops_filter(G):
        g2 = G.copy()
        g2.remove_edges_from(g2.selfloop_edges())
        return g2
    return remove_self_loops_filter
```

B.6.4 Conditions

The enumeration portion of this approach is defined in this `conditionalSubgraphs` function. This allows you to pass in a graph from which you will want to sample subgraphs that meet the conditions that you also pass in.

```
def conditionalSubgraphs(G,condition_list):
    for edges in powerset(G.edges()):
        G_test = G.copy()
        G_test.remove_edges_from(edges)
        if all([c(G_test) for c in condition_list]):
            yield G_test
```

Note that powerset will need to be built (see CBNX for details).

B.6.5 Example condition: requiring complete paths

This condition holds only if a graph has paths from the first node to the second node for each 2-tuple in the node-pair list.

```
def create_path_complete_condition(n_p):
    def path_complete_condition(G):
        return all([nx.has_path(G,x,y) for x,y in n_p])
    return path_complete_condition
```

B.6.6 Non-destructive conditional subgraph generators

Because `conditionalSubgraph` produces an iterator, applying a condition after that initial set is generated, requires splitting it into two copies of the iterator. This involves the `tee` function from the `itertools` core package.

```
def new_conditional_graph_set(graph_set,cond_list):
    graph_set_newer, graph_set_test = tee(graph_set,2)
    def gen():
        for G in graph_set_test:
            G_test = G.copy()
            if all([c(G_test) for c in condition_list]):
                yield G_test
    return graph_set_newer, gen()
```

B.6.6.1 Filters versus Conditions: which to use

The structural differences between filters and conditions highlight how they are to be used. Filters are intended to apply a graph to reduce its edge set in place; as such they return a graph. Conditions return truth values — they are applied to graph set reducing the size of that graph set.

B.7 cbnx: Representing probabilistic relations and sampling

We discuss an algorithm for sampling from Bayesian networks above (sampling). But, most of the difficult parts of encoding a sampling procedure prove (in this case) to do with the al-

gorithm. Rather, the most pressing difficulties arise from attempting to store the relevant information within the NetworkX data dictionaries, so that a self-contained graphical object can be imported and exported. There is a general problem of a lack of standard storage format for Bayesian networks (and probabilistic graphical models in general). This is just one flavor of that problem.

B.7.1 A CBNX IMPLEMENTATION FOR SPRINKLER GRAPH

Below I will illustrate how to use NetworkX[4] and node-associated attributes to define and sample from a parameterized version of the sprinkler Bayesian network represented in abstract, graphical form in Figure B.1. for space reasons comments and formatting were reduced, if you wish to see the original code it can be found at sampling-code.

B.7.2 Sampling infrastructure

```
def sample_from_graph(G,f_dict=None,k = 1):
    if f_dict == None:
        f_dict = {"choice": np.random.choice}
    n_dict = G.nodes(data = True)
    n_ids = np.array(G.nodes())
    n_states = [(n[0],n[1]["state_space"])
        for n in n_dict]
    orphans = [n for n in n_dict
        if n[1]["parents"]==[]]
    s_values = np.empty([len(n_states),k],dtype='U20')
    s_nodes = []
    for n in orphans:
        samp_f = str_to_f(n[1]["sample_function"],
            f_dict)
        s_states = n[1]["state_space"]
        s_dist = n[1]["dist"]
        s_idx = G.nodes().index(n[0])
        s_values[s_idx,:]  = samp_f(s_states,
            size=[1,k],p=s_dist)
        s_nodes.append(n[0])
    while set(s_nodes) < set(G.nodes()):
        nodes_to_sample = has_full_parents(G,s_nodes)
        for n in nodes_to_sample:
            par_indices = [(par,G.nodes().index(par))
                for par in G.node[n]["parents"]]
            par_vals = [(par[0],s_values[par[1],:])
                for par in par_indices]
            samp_index = G.nodes().index(n)
            s_values[samp_index,:] = cond_samp(G,n,
                par_vals,f_dict,k)
            s_nodes.append(n)
    return s_values
```

```
def has_full_parents(G,s_n):
    check_n = [x for x in G.nodes() if x not in s_n]
    nodes_to_be_sampled = []
    for n in G.nodes(data = True):
        if (n[0] in check_n) & (n[1]["parents"]<=s_n):
            nodes_to_be_sampled.append(n[0])
    if len(nodes_to_be_sampled)==0:
        raise RuntimeError("A node must be sampled")
    return nodes_to_be_sampled

def nodeset_query(G,n_set,n_atr=[]):
    if len(n_atr)==0:
        return [n for n in G.nodes(data = True)
            if n[0] in n_set]
    else:
        return_val = []
        for n in G.nodes(data=True):
            if n[0] in node_set:
                return_val.append((n[0],
                {attr:n[1][attr] for attr in n_atr}))
        return return_val

def cond_samp(G,n,par_vals,f_dict, k = 1):
    try: n in G
    except KeyError:
        print("{} is not in graph".format(n))
    output = np.empty(k,dtype="U20")
    for i in np.arange(k):
        val_list = []
        for p in par_vals:
            val_list.append(tuple([p[0],p[1][i]]))
        samp_dist = G.node[n]["dist"][tuple(val_list)]
        samp_f = str_to_f(
            G.node[n]["sample_function"],f_dict)
        samp_states = G.node[n]["state_space"]
        temp_output = samp_f(samp_states,
            size=1,p=samp_dist)
        output[i] = temp_output[0]
    return output
```

```
def str_to_f(f_name, f_dict=None):
    if f_dict == None:
        f_dict = {"choice": np.random.choice}
    try: f_dict[f_name]
    except KeyError:
        print("{} is not defined.".format(f_name))
    return f_dict[f_name]
```

B.7.3 Sampling from the sprinkler Bayes net with cbnx

The following encodes the sprinkler network from Figure B.1 with parameters $p = .2, q_{\text{yes}} = .01, q_{\text{no}} = .4, w_{\text{yes,on}} = .99, w_{\text{yes,off}} = .8, w_{\text{no,on}} = .9 \text{and} w_{\text{no,off}=0}$. This distribution is meant to accord with our intuitions that rain and sprinklers increase the probability of the ground being wet, and that we are less likely to use the sprinkler when it has rained.

```
node_prop_list = [("rain",{
    "state_space":("yes","no"),
    "sample_function": "choice",
    "parents":[],
    "dist":[.2,.8]}),
    ("sprinkler",{
    "state_space":("on","off"),
    "sample_function": "choice",
    "parents":["rain"],
    "dist":{(("rain","yes"),):[.01,.99],
            (("rain","no"),):[.4,.6]}}),
    ("grass_wet",{
    "state_space":("wet","dry"),
    "sample_function": "choice",
    "parents":["rain","sprinkler"],
    "dist":{
        (("rain","yes"),("sprinkler","on")):[.99,.01],
        (("rain","yes"),("sprinkler","off")):[.8,.2],
        (("rain","no"),("sprinkler","on")):[.9,.1],
        (("rain","no"),("sprinkler","off")):[0,1]}})]
```

```
edge_list = [("sprinkler","grass_wet"),
    ("rain","sprinkler"),
    ("rain","grass_wet")]
G = nx.DiGraph()
G.clear()
G.add_edges_from(edge_list)
G.add_nodes_from(node_prop_list)
test = sample_from_graph(G,k=10)
```

B.8 Cognition as Benchmark, Compass, and Map

People have always been able to make judgments that are beyond machine learning's state-of-the-art. In domains like object recognition, we are generally confident in people's judgments as veridical, and – as such – they have been used as a benchmark against which to test and train machine learning systems. The eventual goal is that the system reaches a Turing point — the point at which machine performance and human performance are indistinguishable.

But that is not the only way human behavior can guide machine learning. In domains like causal induction, people's judgments cannot form a benchmark in the traditional sense because we cannot trust people to be "correct". Nonetheless, people *do* make these judgments and, more importantly, these judgments exhibit systematic patterns. This systematicity allows the judgments output by cognition to be modeled using formal, computational frameworks. Further, if we formally characterize both the inputs to *and* outputs from cognition, we can define judgments as optimal according to some model. Formal models of individual cognitive processes can then act as a compass for machine learning, providing a direction for how problems and some solutions can be computed.

Formal frameworks for generating models (e.g., causal theories) can be even more powerful. Data can often be interpreted in multiple ways, with each way requiring a model to generate solutions. Holding the data constant, different goals merit different kinds of solutions. Frameworks that generate models, optimality criteria and solutions not only provide a direction for machine learning, but lay out *sets* of possible directions. Generalized methods that use one system for solving many kinds of problems provide the ability to relate these different directions to each other. Formalizing the inputs, processes and outputs of human cognition produces a map of where machine learning could go, even if it never goes to any particular destination. From this, navigators with more details about the particular terrain can find newer and better routes.

C

A description of the hidden-structure-inference library

This appendix documents the software found in the linked GitHub repository which contains code to perform hidden structure inference over a set of (possibly cyclic) directed graphical models using data expressed in continuous-time. This is the forward sampling method described in subsection 3.12.4 as an alternative to the detective model(subsection 3.12.5) then used in Chapter 3.

This forward sampling is accomplished by using a series of Monte-Carlo approaches to populate the graph parameters and then to sample on the graph. First we engage in metaparameteric sampling, given a distribution on the parameter space, to parameterise the models. Then, given a parameterised graph, it uses Monte-Carlo integration over the joint distribution of observed and hidden events to define a well-formed likelihood for the observed data. This is sufficient to define the marginal likelihood of the observed data given each graph. Then, given a prior distribution over the set of considered graphs, this is sufficient to generate an unnormalised posterior distribution over the set of graphs (given the unnormalised posteriors, normalization is straightforward: divide the unnormalised posteriors by their sum).

C.1 The specific role of this package

The posteriors generated by this system can then be used to calculate the posterior probability that any particular hidden edge in the set of graphs is or is not present. In particular, this software represents an attempt to model the inferences detailed in section 3.12, which models Experiment 1 of Lagnado and Sloman [6], in which the results are expressed in terms of the marginal beliefs about the existence of individual edges under different timing conditions.

Because of this, much of the code in the `likelihood_calculations_shared_parameters.py` will not be easily generalizable to other tasks. In particular, one can only share parameters between graphs if the underlying semantics of the graphs allows that (e.g., the use of superpositioned point processes (as in this case) allows this). Nonetheless, much of the infrastructure will be able to be used in other tasks that involve enumerating and computing functions over sets of graphs.

C.1.1 Graph Enumeration: cbnx+

This uses an enriched version of the Causal Bayesian NetworkX (`cbnx`) library to programmatically enumerate graphical models equipped with rich semantics for operating on graphs based on the semantic or structural properties of the graph's edges and nodes. The basic function used to accomplish this enumeration is `generate_graphs()`, which returns a generator that (for now) needs to be turned into a list to be able to interface with the rest of the code.

The basic graph operations used to specify the set of graphs to be enumerated are `filters` and `conditions`. Given a set of nodes, `cbnx` starts with a complete graph between these nodes, and `filters` take that graph and return a graph with a reduced set of edges according to any rules that will prohibit the existence of an edge in graphs that will be enumerated.

`Conditions`, on the other hand, apply to graphs as a whole to check whether that graph meets or fails to meet a particular condition. Rather than returning a graph, `conditions` return functions that return `Boolean` values that indicate whether a particular enumerated graph did or did not meet that condition.

`registry.py` allows new `filters` and `conditions` to be registered by end-users wishing to adapt this framework to other problems. This is done by importing the `Registry` class and using the `@X.register` decorator before a function, where `X` is replaced by the appropriate class name.

The `filters.py` and `conditions.py` files contain the `filters` and `conditions` needed by

the particular application considered in the hidden-structure-inference package. To register a new `filter`, you can use `@Filter.register`; to register a new `condition` you can use `@Condition.register`.

The node names, particular `filters`, and `conditions` used can be found in `config.py` as the dictionary `generator_dictionary`.

Additionally, it can be faster and more transparent to explicitly list those edges that are to be enumerated over if those are known in advance. These can be included as the optional argument `query_edge_set` that is passed into the `generate_graphs()` function. This feature should be thought of as complementary to `filters` and `conditions`.

C.1.1.1 Graph semantics

Node and Edge properties are currently assigned by applying objects from the `Node_Semantics_Rule` and `Edge_Semantics_Rule` classes found in `node_semantics.py`. Currently, the semantics are defined relative to the naming conventions of the nodes in question. The particular semantics used in the hidden structure inference package can be found in `config.py`, as the dictionaries `edge_semantics` and `node_semantics`.

C.1.2 Graph classes

The continuous-time processes are generated by objects in the `graph_local_classes.py` module.

The `GraphStructure` class encodes graphical structures generated by `networkx` (or other compatible means) in a form that has a more convenient API.

The `GraphParams` class parameterises the edges of the graph encoded in a `GraphStructure` object according to parameter distributions defined by the model. In this case, the `GraphParams` defines parameters for building and sampling a Finitary Poisson Process (a Finitary Poisson Process is a non-homogeneous Poisson Process with a rate function that integrates to a finite value) on each edge of the graph. In particular, this work relies on exponentially decaying rate functions with a maximum rate (⊠) and a decay rate (r), that are defined relative to a scale-free base rate parameter (λ). Multiple edges leading into a node are considered to be superposed on each other (i.e., multiple parent nodes can induce events independently of each other).

The `InnerGraphSimulation` class generates events on these parameterised edges. In this particular application, only the first events generated are relevant to the task, which simplifies

the generation of events.

Caveat: A `InnerNodeSimulation` object is capable of sampling from the more general Finitary Poisson Process, but this needs to be pursued carefully. If no cut-off is given to either the number of generated events or the value the events can take on, this can result in a loop of event generation that will run forever for the purposes of computation. This is true, despite it being the case that at any particular point in the progress of the algorithm, only a finite number of events can be expected from the generated processes. Further details on this are forthcoming.

References

[1] Griffiths, T. L. and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, 116(4).

[2] Greville, W. and Buehner, M. (2007). The influence of temporal distributions on causal induction from tabular data. *Memory & Cognition*, 35:444–453.

[3] Wellen, S. and Danks, D. (2012). Learning Causal Structure through Local Prediction-error Learning. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

[4] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In Gäel Varoquaux, T. V. and Millman, J., editors, *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA.

[5] Pasteur, L. and Chamberland, R. (2002). Summary report of the experiments conducted at pouilly-le-fort, near melun, on the anthrax vaccination, 1881. *The Yale journal of biology and medicine*, 75(1):59. Originally published in Comptes Rendus de l⊠Academie des Science 92:1378-1383, June 13, 1881. Translated by Tina Dasgupta, Yale School of Medicine, Original Contributions Editor, Yale Journal of Biology and Medicine.

[6] Lagnado, D. A. and Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3).

[7] Lagnado, D. A. and Speekenbrink, M. (2010). The influence of delays in real-time causal learning. *The Open Psychology Journal*, 3(2):184–195.

[8] Semmelweis, I. (1983). *The etiology, concept, and prophylaxis of childbed fever*, volume 2. Univ of Wisconsin Press.

[9] Skinner, B. (1956). A case history in scientific method. *American Psychologist*, 11(5):221.

[10] Skinner, B. F. (1964). New methods and new aims in teaching. *New Scientist*, 122(5):483–84.

[11] West-Eberhard, M. J. (2003). *Developmental plasticity and evolution*. Oxford University Press, New York.

[12] Pacer, M., Williams, J., Chen, X., Lombrozo, T., and Griffiths, T. L. (2013). Evaluating computational models of explanation using human judgments. In *Proceedings of the Twenty-ninth Conference on Uncertainty in Artificial Intelligence*. Association for Uncertainty in Artificial Intelligence.

[13] Pacer, M. and Griffiths, T. (2011). A rational model of causal induction with continuous causes. In *Advances in Neural Information Processing Systems*, volume 24, Cambridge, MA. MIT Press.

[14] Pacer, M. and Griffiths, T. (2015). Upsetting the contingency table: Causal induction over sequences of point events. In *Proceedings of the 37th Conference of the CogSci Society*.

[15] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Francisco, CA.

[16] Shimony, S. E. (1991). Explanation, irrelevance and statistical independence. In *Proceedings of the ninth National conference on Artificial intelligence-Volume 1*, pages 482–487.

[17] De Campos, L. M., Gámez, J. A., and Moral, S. (2001). Simplifying explanations in Bayesian belief networks. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(04):461–489.

[18] Flores, M., Gámez, J., and Moral, S. (2005). Abductive inference in Bayesian networks: finding a partition of the explanation space. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 470–470.

[19] Nielsen, U., Pellet, J.-P., and Elisseeff, A. (2008). Explanation trees for causal Bayesian networks. *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*.

[20] Yuan, C. (2009). Some properties of Most Relevant Explanation. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence ExaCt Workshop*, pages 118–126.

[21] Yuan, C., Lim, H., and Lu, T.-C. (2011). Most relevant explanation in Bayesian networks. *Journal of Artificial Intelligence Research*, 42(1):309–352.

[22] Solomonoff, R. J. (1964). A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22.

[23] Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7.

[24] Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24:629–641.

[25] Chater, N. and Vitanyi, P. (2003). Simplicity: a unifying principle in cognitive science. *Trends in Cognitive Science*, 7:19–22.

[26] Baker, A. (2013). Simplicity. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2013 edition.

[27] Feldman, J. (2009). Bayes and the simplicity principle in perception. *Psychological Review*, 116(4):875.

[28] Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press, Cambridge, UK.

[29] Ay, N. and Polani, D. (2008). Information flows in causal networks. *Advances in Complex Systems*, 11(01):17–41.

[30] Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51:354–384.

[31] Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco, CA.

[32] Anderson, J. R. (1990). *The adaptive character of thought*. Erlbaum, Hillsdale, NJ.

[33] Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500.

[34] Shannon, C. E. (1948). The mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

[35] Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3):232–257.

[36] Griffiths, T. L., Lieder, F., and Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229.

[37] Hume, D. (1748). *An enquiry concerning human understanding*. Hackett, Indianapolis, IN.

[38] Mill, J. S. (1843). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence, and methods of scientific investigation*. J. W. Parker, London.

[39] Pavlov, I. (1906). The scientific investigation of the psychical faculties or processes in the higher animals. *Science*, 24(620):613–619.

[40] Ferster, C. B. and Skinner, B. F. (1957). Schedules of reinforcement.

[41] Lacave, C. and Díez, F. J. (2002). A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127.

[42] Yuan, C. and Lu, T.-C. (2007). Finding explanations in Bayesian networks. In *The 18th International Workshop on Principles of Diagnosis*, pages 414–419.

[43] Lombrozo, T. (2012). Explanation and abductive inference. *Oxford handbook of thinking and reasoning*, pages 260–276.

[44] Fitelson, B. (2007). Likelihoodism, Bayesianism, and relational confirmation. *Synthese*, 156(3):473–489.

[45] Tenenbaum, J. B. and Griffiths, T. L. (2001). The rational basis of representativeness. In *Proceedings of the 23rd annual conference of the Cognitive Science Society*, pages 1036–1041.

[46] Abbott, J. T., Heller, K. A., Ghahramani, Z., and Griffiths, T. L. (2012). Testing a Bayesian Measure of Representativeness Using a Large Image Database. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

[47] Sober, E. (2006). Parsimony. In Sarkar, S. and Pfeifer, J., editors, *The Philosophy of Science: An Encyclopedia.*, pages 531–538. Routledge, New York.

[48] Forster, M. and Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1):1–35.

[49] Kelly, K. (2007). How Simplicity Helps You Find the Truth without Pointing at it. *Induction, Algorithmic Learning Theory, and Philosophy*, pages 111–143.

[50] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

[51] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

[52] Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, 103:566–581.

[53] Clark, R. (2001). Information theory, complexity, and linguistic descriptions. In Bertolo, S., editor, *Language Acquisition and Learnability*, pages 126–171. Cambridge University Press.

[54] Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57:227–254.

[55] Monterosso, J., Royzman, E. B., and Schwartz, B. (2005). Explaining away responsibility: Effects of scientific explanation on perceived culpability. *Ethics & Behavior*, 15(2):139–158.

[56] Malle, B. F. (2011). Attribution theories: How people make sense of behavior. pages 72–95. Wiley-Blackwell Malden, MA.

[57] Ahn, W.-K., Novick, L. R., and Kim, N. S. (2003). Understanding behavior makes it more normal. *Psychonomic Bulletin & Review*, 10(3):746–752.

[58] Ahn, W. and Kim, N. (2008). Causal theories of mental disorder concepts. *Psychological Science Agenda*, 22:3–8.

[59] Ahn, W.-k., Proctor, C. C., and Flanagan, E. H. (2009). Mental health clinicians' beliefs about the biological, psychological, and environmental bases of mental disorders. *Cognitive science*, 33(2):147–182.

[60] Frances, A. J. and Egger, H. L. (1999). Whither psychiatric diagnosis. *Australian and New Zealand Journal of Psychiatry*, 33(2):161–165.

[61] Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, 28:107–128.

[62] Read, S. J. and Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, 65(3):429.

[63] Bonawitz, E. B. and Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental psychology*, 48(4):1156.

[64] Lagnado, D. (1994). The psychology of explanation: A Bayesian approach. Master's thesis, University of Birmingham.

[65] Thagard, P. (1989). Explanatory coherence. *Behavioral and brain sciences*, 12(03):435–467.

[66] Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, pages 5–19.

[67] Popper, K. (2002). *The logic of scientific discovery*. Routledge, London & New York.

[68] Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford.

[69] Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, 110(3):499.

[70] Sherman, S. J., Skov, R. B., Hervitz, E. F., and Stock, C. B. (1981). The effects of explaining hypothetical future events: From possibility to probability to actuality and beyond. *Journal of Experimental Social Psychology*, 17(2):142–158.

[71] Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In Kitcher, P. and Salmon, W., editors, *Scientific Explanation*, pages 410–505. University of Minnesota Press, Minneapolis.

[72] Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology*, 52A:273–302.

[73] Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, 8(1):101–118.

[74] Lombrozo, T. and Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99(2):167–204.

[75] Eberhardt, F. and Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of bayesian models. *Minds and Machines*, 21:389–410.

[76] Ai, C. and Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics letters*, 80(1):123–129.

[77] Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61(4):303–332.

[78] Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., and Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115(4):955–84.

[79] Cheng, P. W. and Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, 58:545–567.

[80] Cheng, P. W. and Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, 99:365–382.

[81] Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104:367–405.

[82] Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410.

[83] Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872.

[84] Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. (2010). Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2399–2402. ACM.

[85] Jeffreys, W. H. and Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80(1):64–72.

[86] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

[87] Srebro, N. and Shraibman, A. (2005). Rank, trace-norm and max-norm. In *Proceedings of the Eighteenth Annual Conference on Learning Theory*, pages 545–560.

[88] Crammer, K., Kulesza, A., and Dredze, M. (2009). Adaptive regularization of weight vectors. In *Advances in Neural Information Processing Systems 22*, pages 414–422.

[89] Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–660.

[90] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

[91] Wager, S., Wang, S., and Liang, P. S. (2013). Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems 26*, pages 351–359.

[92] Arriaga, R. I. and Vempala, S. (2006). An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182.

[93] Goodfellow, I. J., Courville, A., and Bengio, Y. (2012). Spike-and-slab sparse coding for unsupervised feature discovery. *arXiv preprint arXiv:1201.3382*.

[94] Powell, D., Merrick, M. A., Lu, H., and Holyoak, K. J. (2016). Causal competition based on generic priors. *Cognitive psychology*, 86:62–86.

[95] Yeung, S. and Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive psychology*, 76:1–29.

[96] Khemlani, S. S., Sussman, A. B., and Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Memory & cognition*, 39(3):527–535.

[97] Schupbach, J. N. (2011). Comparing probabilistic measures of explanatory power. *Philosophy of Science*, 78(5):813–829.

[98] Schupbach, J. N. and Sprenger, J. (2011). The logic of explanatory power*. *Philosophy of Science*, 78(1):105–127.

[99] Williams, J. J. and Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, 34(5):776–806.

[100] Williams, J. J. and Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive psychology*, 66(1):55–84.

[101] Williams, J. J., Lombrozo, T., and Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, 142(4):1006–1014.

[102] Rottman, B. M. and Keil, F. C. (2011). What matters in scientific explanations: Effects of elaboration and content. *Cognition*, 121(3):324–337.

[103] Dahl, R. (1992). The sound machine. In *The Collected Short Stories of Roald Dahl*. Penguin Books Limited, London.

[104] Falcon, A. (2015). Aristotle on causality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2015 edition.

[105] Hume, D. (1739/1978). *A treatise of human nature*. Oxford University Press, Oxford.

[106] Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review*, 57(4):193–216.

[107] Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20(2):158.

[108] Thorndike, E. L. (1911). *Animal Intelligence: Experimental Studies*. Macmillan, New York.

[109] Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical conditioning II: Current research and theory*, pages 64–99. Appleton-Century-Crofts, New York.

[110] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[111] Rumelhart, D. E., McClelland, J. L., Group, P. R., et al. (1986). *Parallel distributed processing*, volume 1. MIT Press, Cambidge, Massachusettes.

[112] Gershman, S. J. (2015). Reinforcement learning and causal models. In Waldmann, M., editor, *Oxford Handbook of Causal Reasoning*. Oxford University Press.

[113] Peirce, C. S. (1883). A theory of probable inference. In Peirce, C. S., editor, *Studies in logic by members of the Johns Hopkins University*, pages 126–181. Little, Brown and Co., New York.

[114] Peirce, C. S. (2014). *Illustrations of the Logic of Science*. Open Court, Chicago, Illinois.

[115] Hacking, I. (1988). Telepathy: Origins of randomization in experimental design. *Isis*, 79(3):427–451.

[116] Fisher, R. A. (1937). *The design of experiments*. Oliver & Boyd, London.

[117] Fisher, R. A. (1992). The arrangement of field experiments. In *Breakthroughs in statistics*, pages 82–91. Springer.

[118] Heider, F. (1958). *The psychology of interpersonal relations*. Wiley, New York.

[119] Kelley, H. H. (1967). Attribution theory in social psychology. In Levine, D., editor, *Nebraska symposium on motivation*, pages 192–238, Lincoln. University of Nebraska Press.

[120] Pearson, K. (1904/1948). On the theory of contingency and its relation to association and normal correlation. In *Karl Pearson's early statistical papers*, pages 443–475. Cambridge University Press, Cambridge.

[121] Wasserman, E. A., Elek, S. M., Chatlosh, D. C., and Baker, A. G. (1993). Rating causal relations: Role of proabbility in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19:174–188.

[122] White, P. A. (2000). Causal judgment from contingency information: The interpretation of factors common to all instances. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:1083–1102.

[123] Buehner, M. and Cheng, P. W. (1997). Causal induction: The Power PC theory versus the Rescorla-Wagner theory. In Shafto, M. and Langley, P., editors, *Proceedings of the*

19th Annual Conference of the Cognitive Science Society, pages 55–61. Lawrence Erlbaum Associates, Hillsdale, NJ.

[124] Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20:557–585.

[125] Gopnik, A., Sobel, D. M., Schulz, L. E., and Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37:620–629.

[126] Sloman, S. (2005). *Causal models: How people think about the world and its alternatives.* Oxford University Press, Oxford.

[127] Lagnado, D. and Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30:856–876.

[128] Feyerabend, P. (1993). *Against method*. Verso.

[129] Pasteur, L. (1880). De l'attenuation du virus du cholera des poules. *CR Acad. Sci. Paris*, 91:673–680.

[130] Koch, R. (1882). *Ueber die Milzbrandimpfung: eine Entgegnung auf den von Pasteur in Genf gehaltenen Vortrag*. Theodor Fischer.

[131] Koch, R. (1987). *Essays of Robert Koch*. New York: Greenwood Press. Carter KC (transl.).

[132] Dolan, E. F. (1958). *Pasteur and the invisible giants*. Dodd, Mead.

[133] Henle, J. (1938). On miasmata and contagie. Translated from the 1840 original by George Rosen.

[134] Evans, A. S. (1993). *Causation and disease: a chronological journey*. Springer Science & Business Media.

[135] Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58(5):295.

[136] Black, B. and Lilienfeld, D. E. (1983). Epidemiologic proof in toxic tort litigation. *Fordham Law Review*, 52:732.

[137] Evans, A. S. (1976). Causation and disease: the Henle-Koch postulates revisited. *The Yale journal of biology and medicine*, 49(2):175.

[138] Buehner, M. J. (2005). Contiguity and covariation in human causal inference. *Learning & Behavior*, 33(2):230–238.

[139] Hammond, L. J. and Paynter, W. E. (1983). Probabilistic contingency theories of animal conditioning: A critical analysis. *Learning and Motivation*, 14(4):527–550.

[140] Kemp, C. (2012). Exploring the conceptual universe. *Psychological Review*, 119(4):685–722.

[141] Qian, T. and Austerweil, J. (2015). Learning additive and substitutive features.

[142] Gopnik, A. and Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, 71:1205–1222.

[143] Schulz, L. and Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, 40:162–176.

[144] Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111:1–31.

[145] Sobel, D. M. and Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, 42:1103–1115.

[146] Kushnir, T., Gopnik, A., Lucas, C., and Schulz, L. (2010). Inferring hidden causal structure. *Cognitive science*, 34(1):148–160.

[147] Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., and Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, 35(8):1407–1455.

[148] Lucas, C. G., Bridgers, S., Griffiths, T. L., and Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, 131(2):284–299.

[149] Mackie, J. L. (1980). *The cement of the universe*. Oxford University Press, London.

[150] Scheines, R., Spirtes, P., Glymour, C., and Meek, C. (1994). Tetrad ii: Tools for causal modeling. *Pittsburgh PA: Carnegie Mellon University*.

[151] Sobel, D. M., Tenenbaum, J. B., and Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, 28:303–333.

[152] Kushnir, T. and Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental psychology*, 43(1):186–196.

[153] Buchsbaum, D., Griffiths, T. L., Plunkett, D., Gopnik, A., and Baldwin, D. (2015). Inferring action structure and causal relationships in continuous sequences of human action. *Cognitive psychology*, 76:30–77.

[154] Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press, New York.

[155] Gallistel, C. R. and Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, 107(2):289.

[156] Pacer, M. and Griffiths, T. (2012). Elements of a rational framework for continuous-time causal induction. In *Proceedings of the 34th Conference of the CogSci Society*.

[157] Hagmayer, Y. and Waldmann, M. R. (2002). How temporal assumptions influence causal judgments. *Memory & Cognition*, 30:1128–1137.

[158] Buehner, M. (2006). A causal power approach to learning with rates. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

[159] Danks, D. (2005). The supposed competition between theories of human causal inference. *Philosophical Psychology*, 18(2):259–272.

[160] Garcia, J., Kimeldorf, D. J., and Koelling, R. A. (1955). Conditioned aversion to saccharin resulting from exposure to gamma radiation. *Science*, pages 1089–1090.

[161] Buehner, M. and May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology Section A*, 56(5):865–890.

[162] Buehner, M. J. and McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, 12(4):353–378.

[163] Krynski, T. R. (2006). *The role of temporal factors and prior knowledge in causal learning and judgment*. PhD thesis, Massachusetts Institute of Technology.

[164] Greville, W. J. and Buehner, M. J. (2010). Temporal predictability facilitates causal learning. *Journal of Experimental Psychology: General*, 139(4):756–771.

[165] Salmon, W. C. (1998). *Causality and explanation*. Oxford University Press Oxford.

[166] Bennett, M. R. (2001). *History of the Synapse*. Harwood, Academic, Amsterdam.

[167] du Bois-Reymond, E. (1848). *Untersuchungen uber tierische Elektrizitat*. Reimer (2 vol.), Berlin.

[168] du Bois-Reymond, P. (1877). Ueber die paradoxen des infinitärcalcüls. *Mathematische Annalen*, 11(2):149–167.

[169] Buckley, B. L. (2012). *The Continuity Debate: Dedekind, Cantor, Du Bois-Reymond, and Peirce on Continuity and Infinitesimals*. Docent Press, Boston, MA.

[170] Grünbaum, A. (1963). *Philosophical Problems of Space and Time*. Alfred A. Knopf, New York.

[171] Van Fraassen, B. C. (1985). *An introduction to the philosophy of time and space*. Columbia University Press, New York, second edition.

[172] Leschiutta, S. (2005). The definition of the 'atomic' second. *Metrologia*, 42(3):S10.

[173] of Weights, I. B. and Measures (2008). *The international system of units (SI)*. US Department of Commerce, Technology Administration, National Institute of Standards and Technology.

[174] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[175] Ross, S. (1983). Stochastic processes. Technical report, John Wiley & Sons.

[176] Kingman, J. F. C. (1993). *Poisson processes*. Wiley Online Library.

[177] Nagaraja, H. N. (2006). Order statistics from independent exponential random variables and the sum of the top order statistics. In *Advances in Distribution Theory, Order Statistics, and Inference*, pages 173–185. Springer.

[178] Dirac, P. A. M. (1981). *The principles of quantum mechanics*. Number 27. Oxford University Press, London.

[179] Eichler, M. and Didelez, V. (2007). Causal Reasoning in Graphical Time Series Models. In *Proceeding of the Twenty-third Conference on Uncertainty in Artificial Intelligence*.

[180] Boyen, X. and Koller, D. (1998). Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth conference on Uncertainty in Artificial Intelligence*, pages 33–42. Morgan Kaufmann Publishers Inc.

[181] Griffiths, T. L. (2005). *Causes, coincidences, and theories*. PhD thesis, Stanford University.

[182] Simma, A., Goldszmidt, M., MacCormick, J., Barham, P., Black, R., Isaacs, R., and Mortier, R. (2008). Ct-nor: representing and reasoning about events in continuous time. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*.

[183] Rottman, B. M. and Ahn, W. (2009). Causal learning about tolerance and sensitization. *Psychonomic Bulletin and Review*, 16(6):1043–1049.

[184] Zhang, J. and Spirtes, P. (2011). Intervention, determinism, and the causal minimality condition. *Synthese*, 182(3):335–347.

[185] Nodelman, U., Shelton, C., and Koller, D. (2002). Continuous time Bayesian networks. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 378–387.

[186] Rajaram, S., Graepel, T., and Herbrich, R. (2005). Poisson-networks: A model for structured point processes. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*. Citeseer.

[187] Salamanca, S. A., Sorrentino, E. E., Nosanchuk, J. D., and Martinez, L. R. (2015). Impact of methamphetamine on infection and immunity. *Frontiers in Neuroscience*, 8:445.

[188] Semmelweis, I. (1988). The etiology, concept, and prophylaxis of childbed fever. (505):46–59.

[189] Hempel, C. G. (1966). *Philosophy of natural science*. Prentice-Hall, New York.

[190] Anderson, J. and Sheu, C.-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition*, 23:510–524.

[191] Chevallier-Jussiau, N. (2009). Henry Toussaint and Louis Pasteur. Rivalry over a vaccine. *Histoire des sciences médicales*, 44(1):55–64.

[192] Macklis, R. M. (1990). Radithor and the era of mild radium therapy. *Journal of the American Medical Association*, 264(5):614–618.

[193] DeVille, K. A. and Steiner, M. E. (1997). The New Jersey Radium Dial Workers and the Dynamics of Occupational Disease Litigation in the Early Twentieth Century. *Missouri Law Review*, 62:281.

[194] Boesch, E. J. and Raber, M. S. (1999). U.S. Radium Corporation. Technical report, Historic American Engineering Record, National Park Service. compiled after 1968.

[195] White, P. A. (2009). Property transmission: An explanatory account of the role of similarity information in causal inference. *Psychological Bulletin*, 135(5):774.

[196] Danks, D., Griffiths, T. L., and Tenenbaum, J. B. (2003). Dynamical causal learning. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances Neural Information Processing Systems 15*, pages 67–74. MIT Press, Cambridge, MA.

[197] Buehner, M. J., Cheng, P. W., and Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29:1119–1140.

[198] Thagard, P. (1998). Explaining disease: Correlations, causes, and mechanisms. *Minds and Machines*, 8(1):61–78.

[199] Schulz, L. E., Bonawitz, E. B., and Griffiths, T. L. (2007). Can being scared cause tummy aches? naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental psychology*, 43(5):1124.

[200] Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47(194).

[201] Bramley, N. R., Gerstenberg, T., and Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

[202] Yeung, S. and Griffiths, T. L. (2011). Estimating human priors on causal strength. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 1709–1714.

[203] Hilton, D. J. and Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1):75–88.

[204] Uttich, K. and Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116(1):87–100.

[205] Bramley, N. R., Lagnado, D. A., and Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(3):708.

[206] Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177.

[207] Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.

[208] Martland, H. S. (1931). The occurrence of malignancy in radio-active persons: A general review of data gathered in the study of the radium dial painters, with special reference to the occurrence of osteogenic sarcoma and the inter-relationship of certain blood diseases. *The American Journal of Cancer*, 15(4):2435–2516.

[209] Lilienfeld, D. E. (2000). John Snow: the first hired gun? *American Journal of Epidemiology*, 152(1):4–9.

[210] Soper, G. A. (1939). The curious career of Typhoid Mary. *Bulletin of the New York Academy of Medicine*, 15(10):698.

[211] Wang, H. and Sun, Y. (2012). An abductive approach to covert interventions. In *Proceedings of the 34th Annual Conf. of the Cognitive Science Society*.

[212] Novick, L. R. and Cheng, P. W. (2004). Assessing interactive causal inference. *Psychological Review*, 111:455–485.

[213] Snow, J. (1855). *On the mode of communication of cholera*. John Churchill, London.

[214] Vinten-Johansen, P., Brody, H., Paneth, N., Rachman, S., Rip, M., and Zuck, D. (2003). *Cholera, chloroform, and the science of medicine: a life of John Snow*. Oxford University Press.

[215] Doyle, A. C. (1894). *The Memoirs of Sherlock Holmes*, volume 2. George Newnes, Limited.

[216] Robbins, H. (1956). An Empirical Bayes Approach to Statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*.

[217] Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII–1983*, pages 1–198. Springer, Berlin.

[218] Pitman, J. (2002). *Combinatorial Stochastic Processes*. Notes for Saint Flour Summer School.

[219] Blei, D. M. and Frazier, P. I. (2011). Distance dependent chinese restaurant processes. *The Journal of Machine Learning Research*, 12:2461–2488.

[220] Teh, Y. W. (2006). A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics.

[221] Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D., Malik, M., Berelowitz, M., Dhillon, A. P., Thomson, M. A., Harvey, P., et al. (1998). Retracted: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, 351(9103):637–641.

[222] Eggertson, L. (2010). Lancet retracts 12-year-old article linking autism to mmr vaccines. *Canadian Medical Association Journal*, 182(4):E199–E200.

[223] Slowik, E. (2014). Descartes' physics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2014 edition.

[224] Nierenberg, D. W., Nordgren, R. E., Chang, M. B., Siegler, R. W., Blayney, M. B., Hochberg, F., Toribara, T. Y., Cernichiari, E., and Clarkson, T. (1998). Delayed cerebellar disease and death after accidental exposure to dimethylmercury. *New England Journal of Medicine*, 338(23):1672–1676.

[225] Wakefield, A., Anthony, A., Murch, S., Thomson, M., Montgomery, S., Davies, S., O'Leary, J., Berelowitz, M., and Walker-Smith, J. (2010). Retraction: Enterocolitis in children with developmental disorders. *The American Journal of Gastroenterology*, 105(5):1214–1214.

[226] Hunter, L., O'Hare, A., Herron, W., Fisher, L., and Jones, G. (2003). Opioid peptides and dipeptidyl peptidase in autism. *Developmental medicine and child neurology*, 45(2):121–128.

[227] Turner, J. R. (2009). Intestinal mucosal barrier function in health and disease. *Nature Reviews Immunology*, 9(11):799–809.

[228] Ritov, I. and Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making*, 3(4):263–277.

[229] Brown, K. F., Kroll, J. S., Hudson, M. J., Ramsay, M., Green, J., Vincent, C. A., Fraser, G., and Sevdalis, N. (2010). Omission bias and vaccine rejection by parents of healthy children: implications for the influenza a/h1n1 vaccination programme. *Vaccine*, 28(25):4181–4185.

[230] Peltola, H., Patja, A., Leinikki, P., Valle, M., Davidkin, I., and Paunio, M. (1998). No evidence for measles, mumps, and rubella vaccine-associated inflammatory bowel disease or autism in a 14-year prospective study. *The Lancet*, 351(9112):1327–1328.

[231] Kaye, J. A., del Mar Melero-Montes, M., and Jick, H. (2001). Mumps, measles, and rubella vaccine and the incidence of autism recorded by general practitioners: a time trend analysis. *British Medical Journal*, 322(7284):460–463.

[232] Farrington, C. P., Miller, E., and Taylor, B. (2001). MMR and autism: further evidence against a causal association. *Vaccine*, 19(27):3632–3635.

[233] Flaherty, D. K. (2011). The vaccine-autism connection: a public health crisis caused by unethical medical practices and fraudulent science. *Annals of Pharmacotherapy*, 45(10):1302–1304.

[234] Rao, T. S., Andrade, C., et al. (2011). The MMR vaccine and autism: Sensation, refutation, retraction, and fraud. *Indian Journal of Psychiatry*, 53(2):95–96.

[235] Smith, M. J., Ellenberg, S. S., Bell, L. M., and Rubin, D. M. (2008). Media coverage of the measles-mumps-rubella vaccine and autism controversy and its relationship to mmr immunization rates in the united states. *Pediatrics*, 121(4):e836–e843.

[236] Jansen, V. A., Stollenwerk, N., Jensen, H. J., Ramsay, M., Edmunds, W., and Rhodes, C. (2003). Measles outbreaks in a population with declining vaccine uptake. *Science*, 301(5634):804–804.

[237] Napier, G., Lee, D., Robertson, C., Lawson, A., and Pollock, K. G. (2016). A model to estimate the impact of changes in MMR vaccine uptake on inequalities in measles susceptibility in Scotland. *Statistical Methods in Medical Research*, 25:1185–1200.

[238] Baker, J. P. (2008). Mercury, vaccines, and autism: one controversy, three histories. *American Journal of Public Health*, 98(2):244–253.

[239] Wilson, G. (1967). *The hazards of immunization.* Athlone Press, London.

[240] Griffiths, T. L., Baraff, E. R., and Tenenbaum, J. B. (2004). Using physical theories to infer hidden causal structure. In Forbus, K., Gentner, D., and Regier, T., editors, *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 446–451. Erlbaum, Mahwah, NJ.

[241] Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 5(2):142–150.

[242] Ghahramani, Z. (1998). Learning dynamic Bayesian networks. In *Adaptive processing of sequences and data structures*, pages 168–197. Springer.

[243] Wingate, D., Goodman, N. D., Roy, D. M., and Tenenbaum, J. B. (2009). The infinite latent events model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 607–614. AUAI Press.

[244] Della Rocca, M. (2005). Two spheres, twenty spheres, and the identity of indiscernibles. *Pacific Philosophical Quarterly*, 86(4):480–492.

[245] Nodelman, U., Shelton, C., and Koller, D. (2003). Learning continuous time bayesian networks. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 451–458.

[246] Gopalratnam, K., Kautz, H., and Weld, D. S. (2005). Extending continuous time bayesian networks. In *Proceedings of the 20th national conference on Artificial intelligence-Volume 2*, pages 981–986. AAAI Press.

[247] Saria, S., Nodelman, U., and Koller, D. (2007). Reasoning at the right time granularity. In *Proceedings of the Twenty-third Conference on Uncertainty in AI*. Best student paper award.

[248] Gatti, E., Luciani, D., and Stella, F. (2012). A continuous time bayesian network model for cardiogenic heart failure. *Flexible Services and Manufacturing Journal*, 24(4):496–515.

[249] Mooij, J. and Heskes, T. (2013). Cyclic causal discovery from continuous equilibrium data. *arXiv preprint arXiv:1309.6849*.

[250] Kan, K. F. and Shelton, C. R. (2008). Solving Structured Continuous-Time Markov Decision Processes. In *International Symposium on Artificial Intelligence and Mathematics*.

[251] El-Hay, T., Friedman, N., Koller, D., and Kupferman, R. (2006). Continuous Time Markov Networks. In *Proceedings of the Twenty-second Conference on Uncertainty in Artificial Intelligence*, Boston, Massachussetts.

[252] Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 438–443.

[253] Hawkes, A. G. (1971b). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

[254] Simma, A. and Jordan, M. (2010). Modeling events with cascades of poisson processes. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*.

[255] Blundell, C., Beck, J., and Heller, K. A. (2012). Modelling reciprocating relationships with Hawkes processes. In *Advances in Neural Information Processing Systems 25*, pages 2600–2608.

[256] Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st national conference on Artificial intelligence-Volume 1*, pages 381–388. AAAI Press.

[257] Guo, F., Blundell, C., Wallach, H., and Heller, K. (2015). The Bayesian Echo Chamber: Modeling Social Influence via Linguistic Accommodation. In *Proceedings of the* 18^{th} *International Conference on Artificial Intelligence and Statistics*, pages 315–323.

[258] Linderman, S. and Adams, R. (2014). Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1413–1421.

[259] Linderman, S. W. and Adams, R. P. (2015). Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*.

[260] Kleinberg, S., Casey, K., and Mishra, B. (2007). Systems biology via redescription and ontologies (i): finding phase changes with applications to malaria temporal data. *Systems and synthetic biology*, 1(4):197–205.

[261] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

[262] Bozdech, Z., Llinás, M., Pulliam, B. L., Wong, E. D., Zhu, J., and DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS Biol*, 1(1):e5.

[263] Einstein, A. (1905). Zur elektrodynamik bewegter körper. *Annalen der physik*, 322(10):891–921.

[264] Galison, P. (2004). *Einstein's clocks and Poincaré's maps: empires of time*. WW Norton & Company.

[265] Ali, M., Goovaerts, P., Nazia, N., Haq, M. Z., Yunus, M., and Emch, M. (2006). Application of poisson kriging to the mapping of cholera and dysentery incidence in an endemic area of bangladesh. *International Journal of Health Geographics*, 5(1):1.

[266] Rathbun, S. L. (1996). Estimation of poisson intensity using partially observed concomitant variables. *Biometrics*, pages 226–242.

[267] Wen, R. and Sinding-Larsen, R. (1997). Stochastic modelling and simulation of small faults by marked point processes and kriging. In Baafi, E. Y., editor, *Geostatistics Wollongong*, volume 1, pages 398–414.

[268] Lloyd, C., Gunter, T., Osborne, M., and Roberts, S. (2015). Variational inference for gaussian process modulated poisson processes. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1814–1822.

[269] Kato, Y., Murakami, Y., Sohmiya, M., and Nishiki, M. (2002). Regulation of human growth hormone secretion and its disorders. *Internal medicine*, 41(1):7–13.

[270] Wingate, D., Kane, J., Wolinsky, M., and Sylvester, Z. (2016). A new approach for conditioning process-based geologic models to well data. *Mathematical Geosciences*, 48(4):371–397.

[271] Nelson, R. (2013). *Probability, stochastic processes, and queueing theory: the mathematics of computer performance modeling*. Springer Science & Business Media, New York.

[272] Killick, D. (2014). Using Evidence From Natural Sciences In Archaeology. In Chapman, R. and Wylie, A., editors, *Material Evidence: Learning from Archaeological Practice*. Routledge.

[273] Hughes, M., Kelly, P., Pilcher, J., and LaMarche, V. (1982). *Climate from tree rings*. Cambridge University Press, Cambridge.

[274] Charlier, P., Poupon, J., Huynh-Charlier, I., Saliège, J.-F., Favier, D., Keyser, C., and Ludes, B. (2009). A gold elixir of youth in the 16th century french court. *British Medical Journal*, 339.

[275] Love, J. (2000). Paleomagnetic principles and practice. *Eos, Transactions American Geophysical Union*, 81(16):172–172.

[276] Greenwood, M. (1921). Galen as an epidemiologist. *Proceedings of the Royal Society of Medicine, Section History of Medicine*, 14:3–16.

[277] Kant, I. (1902). *Prolegomena to any future metaphysics that can qualify as a science*. Open Court Publishing.

[278] Kemp, C., Goodman, N. D., and Tenenbaum, J. B. (2007). Learning causal schemata. In *Proceedings of the Twenth-Ninth Annual Conference of the Cogntive Science Society*.

[279] Cheng, P. W. and Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, 40(1):83–120.

[280] Wolff, P. and Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47(3):276–332.

[281] Sloman, S., Barbey, A. K., and Hotaling, J. M. (2009). A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, 33(1):21–50.

[282] Rottman, B. M. and Keil, F. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, 64(1):93–125.

[283] Rottman, B. (2016). Searching for the Best Cause: Roles of Mechanism Beliefs, Autocorrelation, and Exploitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

[284] Rudy, Y. (2008). Molecular basis of cardiac action potential repolarization. *Annals of the New York Academy of Sciences*, 1123(1):113–118.

[285] Rhodes, D. and Di Luca, M. (2016). Temporal regularity of the environment drives time perception. *PloS one*, 11(7):e0159842.

[286] Newton, I. (1687/1962). *Isaac Newton's Mathematical Principles of Natural Philosophy and his system of the world*. University of California Press, Berkeley and Los Angeles.

[287] Ahn, W.-k., Kalish, C. W., Medin, D. L., and Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54(3):299–352.

[288] Ahn, W.-k. and Kalish, C. (2000). The role of covariation vs. mechanism information in causal attribution. In Wilson, R. and Keil, F., editors, *Cognition and explanation*. MIT Press, Cambridge, MA.

[289] Darwin, C. (1872). *The origin of species*. John Murray, London, 6th edition.

[290] Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

[291] Godfrey-Smith, P. (2013). *Philosophy of biology*. Princeton University Press.

[292] Rutherford, E., Geiger, H., and Bateman, H. (1910). LXXVI. The probability variations in the distribution of α particles. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 20(118):698–707.

[293] Marsden, E. and Barratt, T. (1910). The Probability Distribution of the Time Intervals of a Particles with Application to the Number of a Particles emitted by Uranium. *Proceedings of the Physical Society of London*, 23(1):367–373.

[294] Komar, P., Kessler, E. M., Bishof, M., Jiang, L., Sørensen, A. S., Ye, J., and Lukin, M. D. (2014). A quantum network of clocks. *Nature Physics*, 10(8):582–587.

[295] Grattan-Guinness, I. (1970). *The development of the foundations of mathematical analysis from Euler to Riemann*. MIT Press, Cambridge, MA.

[296] Buchanan, D. W. and Sobel, D. M. (2011). Mechanism-based causal reasoning in young children. *Child development*, 82(6):2053–2066.

[297] Park, J. and Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive psychology*, 67(4):186–216.

[298] Sloman, S. A., Love, B. C., and Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, 22:189–228.

[299] Kim, N. S. and Ahn, W.-k. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology: General*, 131(4):451.

[300] Kim, N. S., Luhmann, C. C., Pierce, M. L., and Ryan, M. M. (2009). The conceptual centrality of causal cycles. *Memory & cognition*, 37(6):744–758.

[301] Rottman, B. M. and Ahn, W.-k. (2011). Effect of grouping of evidence types on learning about interactions between observed and unobserved causes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(6):1432.

[302] Aslin, R. N. (2007). What's in a look? *Developmental Science*, 10(1):48–53.

[303] Carey, S. (2009). *The Origin of Concepts*. Oxford University Press, New York.

[304] Baillargeon, R., Spelke, E. S., and Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, 20(3):191–208.

[305] Bower, T. (1965). The determinants of perceptual unity in infancy. *Psychonomic Science*, 3(1-12):323–324.

[306] Johnson, S., Slaughter, V., and Carey, S. (1998). Whose gaze will infants follow? the elicitation of gaze-following in 12-month-olds. *Developmental Science*, 1(2):233–238.

[307] of Hippo, A. (1860). *Confessions of Augustine*. Warren F. Draper, Andover. Edited, with an introduction by William G. T. Shedd. And, edited according to Act of Congress, in the year 1860, by Warren F. Draper.

[308] Lipton, P. (2003). *Inference to the best explanation*. Routledge, New York.

[309] Kuhn, T. S. (1996). *The structure of scientific revolutions*. University of Chicago Press, Chicago, 3rd edition.

[310] Luce, R. D. (1959). *Individual choice behavior*. John Wiley, New York.

[311] Griffiths, T. L. and Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. *Cognition*, 103:180–226.

[312] Gershman, S. J. and Daw, N. D. (2012). Perception, action and utility: The tangled skein. *Principles of brain dynamics: Global state interactions*, pages 293–312.

[313] Griffiths, T. L., Vul, E., and Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, 21(4):263–268.

[314] Miller, K. T., Griffiths, T. L., and Jordan, M. I. (2008). The phylogenetic Indian Buffet Process: A non-exchangeable nonparameteric prior for latent features. In *Uncertainity in Artificial Intelligence*.

[315] Gershman, S. J., Frazier, P. I., and Blei, D. M. (2015). Distance dependent infinite latent feature models. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):334–345.

[316] Bonawitz, E., Denison, S., Gopnik, A., and Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, 74:35–65.

[317] Abbott, J. T., Griffiths, T. L., et al. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. In *Proceedings of the 34th Conference of the CogSci Society*.

[318] Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press, Oxford.

[319] Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, 126(4):323–348.

[320] Blaisdell, A. P., Sawa, K., Leising, K. J., and Waldmann, M. R. (2006). Causal reasoning in rats. *Science*, 311(5763):1020–1022.

[321] Gershman, S. J., Jones, C. E., Norman, K. A., Monfils, M.-H., and Niv, Y. (2013). Gradual extinction prevents the return of fear: implications for the discovery of state. *Frontiers in behavioral neuroscience*, 7:164.

[322] Suchow, J. W., Pacer, M. D., and Griffiths, T. L. (2016a). Design from zeroth principles.

[323] Suchow, J. W., Morgan, T. J. H., Hamrick, J., Pacer, M., Meylan, S. C., and Griffiths, T. L. (2016b). Wallace: automating cultural evolution experiments through crowdsourcing. Oral presentation.

[324] Gallistel, C., Mark, T. A., King, A. P., and Latham, P. (2001). The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *Journal of Experimental Psychology: Animal Behavior Processes*, 27(4):354.

[325] Pacer, M. D. (2015). Causal-Bayesian-NetworkX. http://dx.doi.org/10.6084/m9.figshare.1471763. Online (accessed September 2, 2015).

[326] McKay, B. D., Oggier, F. E., Royle, G. F., Sloane, N., Wanless, I. M., and Wilf, H. S. (2004). Acyclic digraphs and eigenvalues of (0, 1)-matrices. *Journal of Integer Sequences*, 7:1–5.

[327] Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30.

[328] Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. http://www.scipy.org/. Online (accessed 2015-07-04).

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Adobe Minion Pro, a transitional humanist typeface inspired by the old style typefaces of the late Renaissance. It has been set according to the rules laid out by the Berkeley Graduate Division. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online on github at github.com/suchow/Dissertate, at Dissertate, from Dissertate's lead author, Jordan Suchow, at suchow@post.harvard.edu, or the developer of the template, Michael Pacer, at mpacer@berkeley.edu