Edge Replacement as a Model of Causal Reasoning

By David W. Buchanan

B. A., McGill University, 2003

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the Department of Cognitive, Linguistic and Psychological Sciences at Brown University

Providence, Rhode Island

May 2011

This dissertation by David W. Buchanan is accepted in its present form by the Department of Cognitive, Linguistic and Psychological Sciences as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Dr. David Sobel, Advisor

Date _____ _____

Dr. Joshua Tenenbaum, Reader

Date _____ _____

Dr. Kathryn Spoehr, Reader

Date _____ _____

Dr. Thomas Serre, Reader

Approved by the Graduate Council

Date _____ _____

Dr. Peter Weber, Dean of the Graduate School

Curriculum Vitae

David W. Buchanan

Department of Cognitive, Linguistic, and Psychological Sciences
Brown University, Box 1821
Providence, RI 02912
David_Buchanan@brown.edu

Born: July 11, 1978, Montreal, Canada

Research Interests

Causal reasoning in children and adults, particularly reasoning about mechanisms, Computational models of cognitive development.

Education

2003	B.A. with distinction, McGill University.
2011 (Expected)	Ph.D., Cognitive Science, Brown University.

Peer-Reviewed Publications

Buchanan, D. W., & Sobel, D. M. (2008). Children's developing inferences about object labels and insides from causality-at-a-distance. *Proceedings of the 30th annual meeting of the Cognitive Science Society*.

Sobel, D. M., & Buchanan, D. W. (2009). Bridging the gap: Causality at a distance in children's categorization and inferences about internal properties. *Cognitive Development*, 24, 274-283.

Buchanan, D. W., & Sobel, D. M., (2010). Causal stream location effects in preschoolers. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.*

Buchanan, D. W., Tenenbaum, J. B., & Sobel, D. M. (2010). Edge replacement and nonindependence in causation. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Buchanan, D. W., & Sobel, D. M. (in press). Mechanism-based reasoning in young children. *Child Development.*

Sobel, D. M., Buchanan, D. W., Butterfield, J., & Jenkins, O. C. (in press). Interactions between models, theories, and social cognitive development. *Neural Networks.*

Sobel, D. M., Sedivy, J., Buchanan, D. W., & Hennessey, R. (in press). The role of speaker reliability in judgments about mutual exclusivity. *Journal of Child Language*.

Buchanan, D.W., & Sobel, D.M., (under review, *Cognitive Science*). Causal stream location effects in preschoolers. Brown University.

Other Conference Presentations

Buchanan, D. W., & Sobel, D. M. (2011, April). *Complexity and Determinism in Preschoolers' Causal Reasoning* Talk presented at the 2011 Biennial meeting of the Society for Research in Child Development, Montreal, Canada.

Buchanan, D. W., & Sobel, D. M. (2009, April). *Children's developing causal inferences from mechanism and covariation information*. Poster presented at the 2009 Biennial meeting of the Society for Research in Child Development, Denver, CO.

Awards and Honors

2009 *Graduate Student Research Award* (Brown University, Brain Sciences Program), $9,813 over one year.

2009 *Society for Research in Child Development Student Travel Award* $300 for travel and related expenses, to attend and present at biannual conference in Denver, Colorado.

2007 *Graduate Summer School: Probabilistic Models of Cognition: The Mathematics of Mind* (Institute for Pure and Applied Mathematics, University of California at Los Angeles) $500 for travel, plus to room and board for 3 weeks, to attend this summer school, provided by the National Science Foundation.

Teaching

2008-2009: Completed Teaching Certificate Level I, at the Sheridan Center for Teaching and Learning, Brown University.

Fall 2008: Teaching Assistant for COGS 0090: *Quantitative Methods in Psychology*, Brown University. Instructor: Fulvio Domini.

Spring, 2009: Teaching Assistant for COGS 0630: *Children's thinking: the nature of cognitive development*. Brown University. Instructor: David Sobel.

Fall, 2009: Teaching Assistant for COGS 0630: *Children's thinking: the nature of cognitive development*. Brown University. Instructor: James Morgan.

Spring, 2010: Teaching Assistant for COGS 1280: *Computational Cognitive Science*, Brown University. Instructor: Thomas Serre.

PREFACE AND ACKNOWLEGDEMENTS

This thesis would not have been possible without the help of my colleagues, friends, and family. My advisor David Sobel provided unwavering support, guidance, and feedback over five years of graduate school, contributing in particular to the theory and experiments described in this thesis. Joshua Tenenbaum was instrumental in guiding and helping develop the edge replacement model. Other members of my thesis committee, Thomas Serre and Kathryn Spoehr, provided invaluable help along the way. Thanks also to Steven Sloman for work on my preliminary examination, which formed much of the basis for the work here. I also wish to thank my colleagues Naomi Feldman and Adam Darlow for helping with some of the mathematics, Phil Fernbach for his insight into the theoretical and philosophical aspects of causality, and Hugh Rabagliati and Deanna Marcis for help with experimental design. I also wish to thank the research assistants at the Causality and Mind Lab at Brown university for help with data analysis, in particular Katie Green, Aria Auerbach, and Marisha Gadowski. Noah Goodman and Thomas Griffiths helped by providing feedback about mathematical aspects of the model, and Ralf Mayrhofer and Bob Rehder generously shared details of their experimental data. Thanks to my mother, Carrie Buchanan, my father, George Buchanan, my mother-in-law Joanne Mitchell, my father-in-law Doug Weed, for their support and encouragement. Thanks most of all to my wife Sarah Mitchell-Weed, for her love, poetry, cooking, and proofreading.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

"To the synthesis of cause and effect there belongs a dignity which cannot be empirically expressed, namely that the effect not only succeeds upon the cause, but that it is posited through it and arises out of it."

-Immanuel Kant

"I have not been able to discover the cause of those properties of gravity from phenomena, and I frame no hypotheses; for whatever is not deduced from the phenomena is to be called a hypothesis, and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy."

-Isaac Newton

Human beings come to learn about the causal structure of the world. Think of an advanced adult state, such as an auto mechanic or doctor. These adults have complex, powerful representations of the causal structures in the domains of their expertise. When your car fails to start, the auto mechanic can tell you whether the problem is a faulty starter solenoid, or a problem with the fuel injection system. A doctor can reason effectively about which novel treatments are likely to prevent disease. Even outside our expertise, we know a great deal about the causal structure of the world around us. Even if you do not know the details of how a helicopter works (Keil, 2003), you know that it is not likely to work underwater, that like a car, it probably needs some kind of fuel, but unlike a car, it probably does not have brakes. It is worthwhile to reflect on the sophistication of our causal reasoning, even in these situations where we have limited specific knowledge. How we reason so flexibly and effectively is still largely a mystery.

A further mystery is how such knowledge develops. While some causal reasoning abilities may be innate, it is unlikely that reasoning about mechanical and electronic artifacts is specified in our genes. Such knowledge would have had to evolve, but evolution has not had enough time to specialize humans to deal with electricity, let alone television remotes and cell phones. Rather, it is likely that the brain has a set of learning processes that allow us to develop the complex representations of the auto mechanic, doctor, or reasonable novice in helicopter engineering, from earlier (probably simpler) representations of the child, together with data from the world. Given the dearth of doctors and engineers in our society, the benefits of understanding this process should be obvious.

This thesis will not present a complete solution to how such learning occurs. Such a solution can only result from the efforts of a broad research program, involving many researchers and many perspectives and methodologies. Instead, it will focus specifically on a subpart of this question. Much of our adult knowledge involves knowledge not just of causes and effects, but of what happens *between* cause and effect. For instance, we all know that turning the key causes the car to start, but the auto mechanic has detailed knowledge about which intermediate events enable this relation. When and how do we come to represent such intermediate events? This is related to the long standing question of mechanism.

A focusing question: mechanism

There are a wide variety of theories about how we develop representations of cause and effect, which can be differentiated in a wide variety of ways. But arguably the dimension on which they disagree the most is whether causal representations necessarily involve representations of mechanisms: the events that occur between cause and effect. Some theories, for instance, argue that causal representations capture purely statistical, (Cheng, 1997) or counterfactual (Lewis, 1973) properties of relations between events. To these theories, a causal relation between A and B means no more and no less than A's capacity to bring about B. To other theories, the causal relation between A and B necessarily involves something more: A representation of *how* A brings about B. In this camp, we can include theories of generative transmission (Shultz, 1982) and force dynamics (Wolff, 2007).

As an example, imagine researchers discovered (through replicated, double blind, placebo controlled studies), that eating blue jelly beans caused people to lose weight. This is not just correlation but real causation: we can reliably bring about weight loss by making people eat the jelly beans. Under a 'mechanism-free' account, a person should be able to represent this relation directly: there is a cause, an effect, and a demonstrable relation between the two events that obeys the appropriate statistical rules. Under a 'mechanism-dependent' account, a person would additionally have to imagine some mechanism (possibly vague) that connected the two events, as a necessary part of representing the causal relation. Anecdotally, people seem to find

either the mechanism-based or mechanism-free account intuitively right, and are stunned that another person could hold the other view. Disagreements abound, some productive, and some not. To some of us, (like Kant) such an unexplained statistical relation is metaphysically irritating. To others, (like Newton), such questions are distraction from real science, whose job is to discover laws, even if the mechanism is unknown.

Some disagreements about concepts are merely philosophical – often two theories predict and explain the same phenomena. This is not one of those cases. Different answers to the question of mechanism make large differences to the phenomena that theories of causal representations can predict and account for. And yet, the question is still far from settled. There are two reasons for this. One is that the empirical, psychological question of mechanism, is bound up with, and often confused with, a larger normative question. Medical researchers, for instance, argue about whether a mere statistical relation, however powerful, should be sufficient to infer causation before we understand the mechanism (e.g. Weed & Hursting, 1998). While such debates are psychologically interesting and sometimes relevant, (even some trained researchers still want to know 'how', not just 'why') this thesis will not try to resolve them. The question is not whether people *should* represent mechanism as an essential part of causation, but whether they *do.*

The second reason that the debate has not been resolved, is that there are not equivalent, comparable formal models on each side. The last fifteen years in cognitive

science have seen the emergence of a powerful set of formal tools for doing causal inference, known as Causal Graphical Models (Pearl, 2000/2009; Spirtes, Glymour, & Scheines, 1993/2000), which this thesis will sometimes call CGMs. These models, as currently employed, amount to a well-worked out version of the mechanism-free account of human causal representations.[1] They make clear and specific predictions about how human beings should behave with respect to causal systems – some correct, and some (this thesis will argue) incorrect. Unfortunately, similarly rigorous models from the mechanism camp (e.g. Wolff, 2007) do not apply to nearly the same range of phenomena. Furthermore, the relative vagueness of most theories on the mechanism side makes it difficult for them to make falsifiable predictions. It is therefore difficult to empirically compare the two accounts.

This thesis will hopefully take a small step toward addressing this imbalance. The central argument is that we can amend Causal Graphical Models in such a way that they support a more accurate, and also more mechanism-dependent account of causal representations. CGMs currently use a *minimality* constraint: strictly prefer models that posit fewer causal relations. It is this commitment that carries the mechanism-free commitment of CGMs as they are currently employed. This thesis will argue that we can replace minimality with a novel edge replacement rule, removing this commitment and importing at least some aspects of mechanism into CGMs. The edge replacement rule

[1] Not all researchers who have developed causal models intend them as part of a mechanism-free account. Pearl (2000), for instance, explicitly endorses aspects of the mechanism-based account. We will see that Causal Graphical Models, as currently applied, do implicitly carry many of the current commitments of the mechanism-free account.

makes new, correct predictions, in many places where minimality makes incorrect predictions. It also makes several novel predictions, which we will begin to test in the later chapters. Unfortunately, the specificity of these predictions means we must leave behind the comfortable vagueness of our original mechanism-based intuitions. The model is specific, and therefore will almost certainly be proved wrong in the long run. If so, it will likely be replaced by a better model. Many people will not consider this an account of mechanism at all; there is certainly more work to do. It is at least worthwhile to begin.

Overview

This section presents an overview of the content of each chapter.

Chapter 1 (the next chapter) will present Causal Graphical Models as a formal and computational framework in which to address the question of mechanism. Causal Graphical Models currently use the principle of *minimality*, a strong preference for simplicity in the graphs that are posited to explain a given causal relation. Chapter 1 will discuss a series of problems that arise from the use of minimality: That minimality often causes mechanism to be left out of the theoretical picture, that minimality cannot explain people's apparent preference for determinism, and that minimality cannot explain why people expect collateral effects to be correlated given a common cause (a *nonindependence effect*).

Chapter 2 will present an alternative to minimality: A generative edge replacement process through which we can construct Causal Graphical Models. This is the main novel contribution of the thesis. Edge replacement assigns a probability to each graph, depending on how likely it is to be generated according to recursive application of an edge replacement rule. Chapter 2 also discusses extensions to edge replacement, including the introduction of time, and a method for doing inference over classes of graphs that have relevantly similar structure. This chapter also describes how edge replacement explains nonindependence effects and determinism, and discusses edge replacement's implications for understanding representations of causal mechanisms. This discussion gives rise to a set of empirical predictions that will be tested experimentally in Chapters 3, 4, and 5.

Chapter 3 investigates whether preschoolers (3- and 4-year-olds) will show a phenomenon called a *stream location effect*, as edge replacement predicts. Preschoolers are chosen because their representations are likely to be more basic than those of adults. For this reason, edge replacement and minimality make the most divergent predictions about the representations of preschoolers, particularly concerning novel causal systems. Stream location effects arise from the branching character of edge replacement, but not from the simpler graphs preferred by minimality. Chapter 3 shows that preschoolers show a robust stream location effect, even in a novel causal system, and even for a pattern of data that is contrary to their existing mechanism knowledge.

Chapter 4 tests the temporal component of edge replacement; in particular, the commitment that variability arises not from randomness (as minimality would predict) but from the presence of hidden inhibitors that tend to persist in time. Experimental results show that 4-year-olds reason in accord with the predictions of edge replacement, and contrary to the predictions of minimality. 3-year-olds show chance performance, consistent with either model.

Chapter 5 brings the focus of the thesis back to mechanism. We can see an unknown causal mechanism connecting cause and effect as a metaphorical 'black box,' whose internal structure is unknown. Edge replacement and minimality formalize different commitments about the 'contents' of the box. In particular, edge replacement implies that given a variable causal relation, the physical structure of the mechanism is more likely to be complex, than given a simple causal relation. Chapter 5 describes experiments designed to test whether preschoolers recognize this principle: Variability implies complexity. Results support the hypothesis that 4-year-olds understand this principle.

The conclusion will begin by recapping the results of the model and experiments. It will then discuss further experimental work, in both adults and children. This will be followed by a discussion of the prospects for further developing the model: using edge

replacement as a learning mechanism for elaborating representations of mechanism, and discovering an underlying process through which edge replacement might itself be learned. Finally, the conclusion will argue that Causal Graphical Models as a amended by edge replacement can be seen as an alternative rational model of causal reasoning. That is, edge replacement is designed to describe optimal reasoning, not deviations from the 'correct' reasoning as described by minimality.

CHAPTER 1: CAUSAL GRAPHICAL MODELS AND MINIMALITY

This thesis will take as a starting point the framework of Causal Graphical Models. While there are other approaches to modeling causal reasoning, CGMs were chosen for a few important reasons. First, they are arguably the most powerful and expressive framework currently available, for representing causal reasoning and learning. Several other models of causal reasoning, for instance, can be seen as special cases of Causal Graphical Models. An example is Cheng's (1997) Power PC model, which Glymour (1998) showed could be recovered formally from CGMs. The predictions made by CGMs depend to a great extent on the auxiliary assumptions made; for this reason, this section is best seen as outlining a framework for comparing models, rather than a theory or model in itself. By situating multiple models within this same framework, we are better able to make a direct comparison between models. The power of CGMs ensures that multiple models can speak the same language, avoiding ambiguity.

Another reason to prefer CGMs is that they currently form the basis for some of the most active research programs in the study of causal reasoning. As we will see below, they show the greatest promise for addressing the problem that raised in the introduction: How is it that humans develop complex representations of causal structure?

An example

Before proceeding, it is necessary to set up an example that we will return to throughout the thesis. In the long tradition of models of causal reasoning (e.g. Pearl, 1988), the example involves a car. Imagine you live in a cold climate, which for simplicity's sake has two temperatures: Cold, and Very Cold. Every day, you need to start your car in the morning. You walk out into the snow, and turn the key in the ignition, hoping that the car will start. About half the time it starts, and about half the time it does not. Other things also happen when you turn the key: for instance, most of the time your lights dim when the engine is turning over, regardless of whether the car actually starts. When your car fails to start, you need to ask your neighbor for a ride, (for some reason, his car always seems to start). One day, you notice something: Almost every day that your car fails to start, is a day on which it is Very Cold. From this point on, on nights when the forecast is for Very Cold, you call your neighbor in advance, and arrange a ride to work. Your boss is happier now, because you show up to work on time more often. But there is still a lingering issue: There are some Very Cold days on which your car does start. You then notice that every Very Cold day, on which your car does start, is a day on which you had gotten up in the middle of the night to drive to a local store and get a snack. You reason that the drive to the store must be acting somehow on the temperature of the engine in the morning. You start waking up every Very Cold night, hungry or not, and running the engine for a few minutes. This solves your

problem: Your car starts every morning from then on. This makes both your boss, and your neighbor, much happier. You consider getting a heated garage like your neighbor.

This is an example of coming to make a more complex causal representation out of a simple one, and thus is a good source of intuitions about our focusing question. The complexity of this example pales in comparison to the complexity faced by the auto mechanic, medical student, or even a preschooler learning about the world. It is just complex enough to challenge CGMs towards considering the right types of learning mechanisms for answering our focusing question of how can we come to learn about complex causal structures.

Strength and Structure

Some models of causal reasoning and learning focus exclusively on learning the strength of a causal relation. A classic example is the Rescorla-Wagner learning rule, (Rescorla & Wagner, 1972) which provides an account of how animals should change their behavior trial by trial, as they are exposed to data about causes and effects. Most of these models hold at their core some calculation of the proportion of successes and failures of causes, in producing their effect. For instance, given an effect that occurs 50 per cent of the time in the presence of the cause, that never occurs in the absence of the cause, most strength based models will infer, or converge to, a strength of 0.5. Many strength-based models have been successful in predicting how humans and nonhuman animals will change their behavior trial by trial. Ultimately, these models are tested by how well they predict some behavioral indicator of represented strength (such

as lever or button presses, or probability judgments) given some set of contingencies, often trial-by-trial.

As sound as many of these models are, they are not suited to answer the question at hand: how we come to have complex representations of causal structure. This is because structure cannot be reduced to strength. Strength is a quantitative construct, but structural questions are often qualitative: For instance: often you want to know whether a causal relation exists, what intermediate events exist, the direction of the causal relation, and what interventions would change the causal relation. As an example, imagine you brought your car to your mechanic, complaining that it only started half the time. You ask the question: "What is wrong with it?" A strength based model would have to respond: "The probability of your car starting is 0.5." But this is a non-answer – if your mechanic said this, you would find a new mechanic. You'd much rather hear an account that involved the mechanical or electrical systems. This is not just a matter of language or semantics: The auto mechanic is capable of answering questions about both strength ("What is the probability of my car starting?") and structure ("What is wrong with my car?"). A strength-based model does not have structural information, even when the question is clarified. Only a model that includes inferences about causal structure shows the potential to answer structure-based questions properly.

There are ways to twist strength-based models into answering structural questions. For instance, if the question we want answered is *whether* a causal relation

exists, we could use strength of association as a proxy for the probability that a causal relation exists. This approach, however, turns out to yield incorrect predictions about human behavior in many cases (e.g. Griffiths & Tenenbaum, 2005; Sobel, Tenenbaum, & Gopnik, 2003). Attempts to deflate structure to strength often fail. Another solution is to give the strength-based model a great deal of help. For instance, by feeding the model extensive data about with cars, weather, and garages, such a model might output something sensible as a mechanic. It might say, for instance, that the probability of a starting problem is high, given cold. But such a tactic begs all the important questions. How did the model represent that temperature was relevant? How is a specific strength judgment developed and retrieved from a whole class of possible enablers, disablers, and causes? And how can we use a strength-based model to recognize what changes are likely to affect a causal system, as we did when we started the car in the middle of the night? These questions are structural, and thus external to a strength-based model. To answer them with any kind of detail or power, we must look beyond strength.

As models of the phenomena for which they are designed, strength-based models are good science. As applied to the specifically structural questions raised in the introduction, they are either radically (and incorrectly) deflationary, or leave too much out. Instead, we will use a framework that allows for both strength and structure: Causal Graphical Models.

Nodes, edges, and functional forms

This section will present a formal overview of the Causal Graphical Models framework. The basic principles are drawn from foundational work by Spirtes, Glymour, and Scheines, (1993/2000) as well as Pearl (2000/2009), building on ideas that can be traced as far back as Reichenbach (1956). They are also drawn from developments attributable to many researchers too numerous to mention exhaustively (e.g. Gopnik, Glymour, Sobel, Schulz, Kushnir & Danks, 2004; Griffiths & Tenenbaum, 2005; Sloman, 2005). Many of these latter developments concern the application of CGMs specifically to psychology. This section will outline the formal principles of CGMs in the broadest possible terms, so that it can serve our main purpose: Situating and comparing multiple theories of causal representation. As such, these principles are not original work, but they do bear a specific interpretation, emphasis and choice of terminology, that will be necessary to facilitate the introduction of original work in Chapter 2.

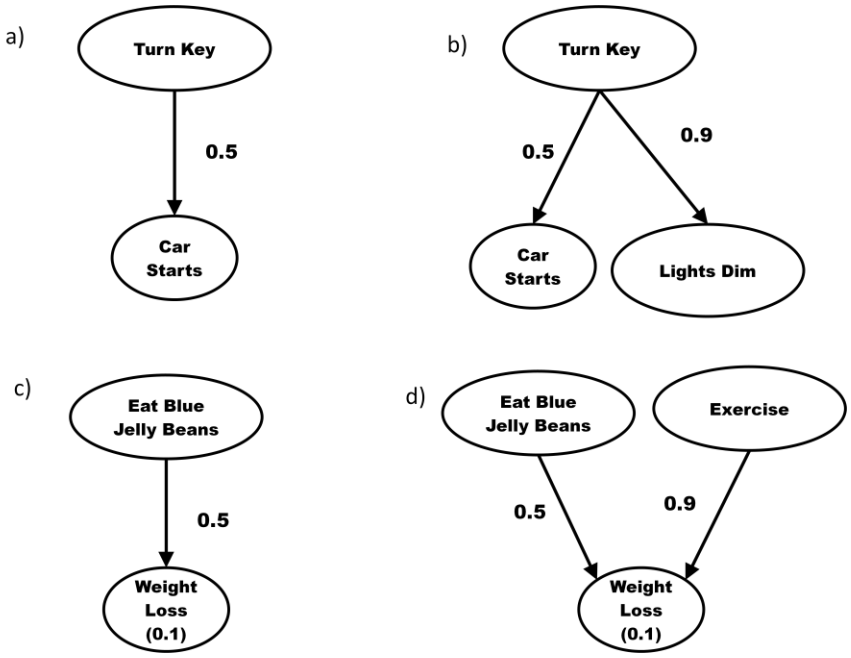

Figure 1: A series of CGMs

Under Causal Graphical Models, a system of causal relations is represented by a graph, made up of nodes and directed edges. Nodes represent events, and edges represent causal relations. Each edge has a direction that specifies the direction of causation. Each node also has a *functional form* -- a set of rules that capture the causal relations in which it participates. The rules specify exactly what happens to the effect, given each possible combination of causes. Together, the nodes, edges, and functional forms, allow us to determine the status of events given other events. An example is given in Figure 1a: the edge connecting the two nodes indicates that turning your key causes your car to start.

There is also a functional form to this relation, which remains to be specified. Formally, CGMs can employ a wide variety of functional forms. In practice, and especially in psychology, a few specific functional forms are used by default. This section will introduce a few of these (simple generative, preventative, and Noisy-OR for binary variables) along the way to outlining the rules of CGMs. In this case, the functional form is that turning the key in your car's ignition system causes the car to start, with probability 0.5. Functional forms can be expressed unambiguously using a conditional probability table like the one below, which specifies the rules for 'car starts.'

	Car Starts=Yes	Car Starts=No
Turn Key= Yes	0.5 (p)	0.5 (1-p)
Turn Key= No	0	1

This table specifies a probabilistic causal relation. Given that we turn the key, the probability of the car starting is 0.5. Given that we do not turn the key, the probability of the car starting is 0. Throughout this thesis, this type of relation will be illustrated graphically by writing a number between zero and one beside the edge, as in Figure 1. The number is shorthand for the conditional probability table shown above: If 'p' is the number shown over the edge, then the probability of the effect given the cause is p, the probability of not having the effect given the cause is 1-p, the probability of the effect without the cause is zero, and the probability of not having the effect without the cause is 1. Edges without an attached number will always be deterministic; that is, p=1.

So far, we have used Causal Graphical Models to create a simple strength-based model: By plugging in a simple learning algorithm for 'p,' we could recover a simple proportional strength-based model. But CGMs have far more expressive power than this simple model (as do many more sophisticated strength-based models). One example is their ability to model additional effects. For instance, 'engine starts' is not the only thing that happens when you turn the key: The lights also dim most of the time. CGMs let you accommodate this event by adding an additional event, as in Figure 1b. We have added a causal relation between turning the key, and the lights dimming. This means, now, that when we turn the key, there is a 90 per cent chance that the lights will dim, and a 50 per cent chance that the engine will start. This type of causal structure is known as a *common cause* structure (Reichenbach, 1956/1991): one event causes more than one other event.

Common effects are also possible: an event can have more than one cause. There are two ways of handling this in CGMs. One is to say that the event sometimes occurs because of the cause, and sometimes occurs due to factors that are not captured in the graph. From the point of view of the graph, this is just like the event occurred spontaneously. We will use a lowercase 'a' to refer to this probability of spontaneous 'activation.' The car example is not suited to this case, because cars almost never start without the owner turning the key. But it often occurs in medical scenarios, where disease or recovery can happen independently of the treatment being tested. Figure 1c shows a CGM of the jelly bean example from the introduction. In this case, the

probability that eating jelly beans will cause a person to lose weight is p=0.5, and the probability that they will spontaneously lose weight is a=0.1. We can combine these probabilities using a 'Noisy-OR' parameterization as shown in this table:

	Lose Weight=Yes	Lose Weight=No
Eat Jelly Beans= Yes	p+a-pa=0.55	1-(p+a-pa)=0.45
Eat Jelly Beans = No	a=0.1	(1-a)=0.9

The Noisy-OR parameterization was originally applied to causation by Cheng (1997). Glymour (1998) has shown that a Bayesian analysis of common effect situations, using a Noisy-OR parameterization, can recover Cheng's strength-based model. A Noisy-OR captures the intuition that multiple causes can each independently bring about the effect. It is straightforward to generalize the Noisy-OR pattern to multiple causes. Throughout this thesis, any graph which contains multiple generative edges that end in the same node, will be assumed by default to follow a Noisy-OR functional form, unless otherwise specified.

Finally, there are preventers. For instance, in the car example, it might be that very cold weather prevents your car from starting. An example of this is shown in Figure 2. In this case, very cold weather, which occurs about half the time, always disables the relation between 'turn key' and car starts. Formally, disabling uses the functional form that the effect occurs if and only if both the cause occurs, and the preventer does not. A table is shown below. Throughout the thesis, we will indicate such a relation using a dashed edge.

	Car Starts=Yes	Car Starts=No
Turn Key= Yes, Cold Weather=Yes	0	1
Turn Key= Yes, Cold Weather=No	1	0
Turn Key= No, Cold Weather=Yes	0	1
Turn Key= No, Cold Weather=No	0	1

Other functional forms are possible; for instance, AND functions, and more complex combinations of multiple variables. As long as row probabilities sum to one, arbitrary conditional probability tables are possible. We are not even restricted to discrete variables – a functional form could be a function mapping one continuous variable to another. For instance, weight can be seen as a function of height plus other factors or random noise. For the sake of simplicity, this thesis will not deal with continuous variables, but they are part of the wide scope of CGMs.

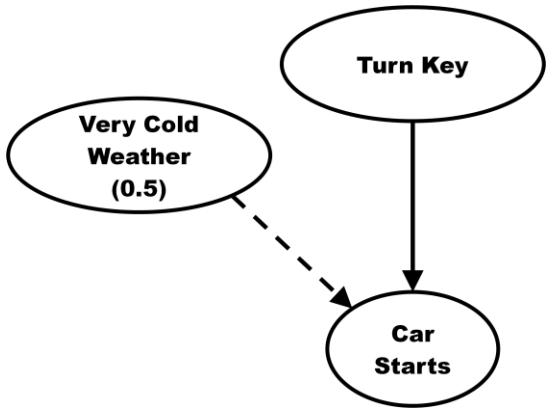

Figure 2: A prevention relation

The real power of CGMs comes not from looking at these simple examples (which don't go much beyond strength-based models on their own) but from combining

them into a more complex graph. Figure 3 shows one CGM that captures the causal system from the car example.

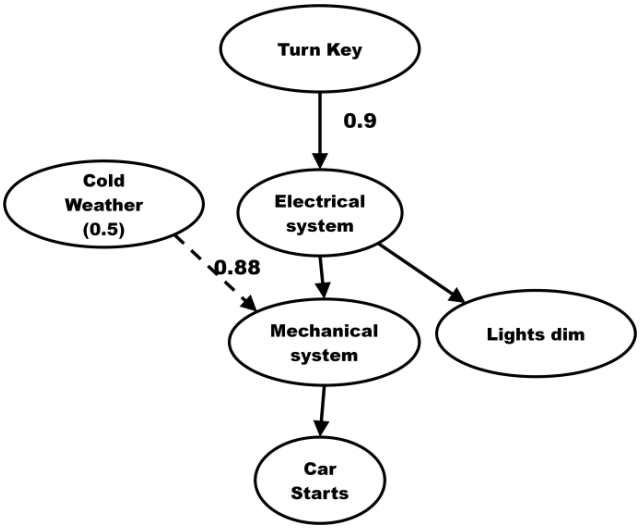

Figure 3: A representation of the car example.

If this is our representation of the causal structure of the car, then it implies the following set of properties: Whenever we turn the key, there is a 90 per cent probability that the electrical system engages. This always dims the lights. The electrical system engages the mechanical system, which can be disabled by cold weather. As long as the mechanical system is so engaged, the car will start. By doing the math over the series of functional forms, we can see that the probability that the car starts given that we turn the key is (1-0.5*0.88)*0.9=0.504. We can also deduce more complex dependence relations. If any of these relations fail to hold, then it implies that we need to amend our representation to better fit the underlying causal structure of the world. Having outlined some of the basics of CGMs, we will now discuss some formal properties that arise from them.

The Markov condition

Implicit in the definition of CGMs is the Markov condition: Events depend only on their direct parents. This is built into the most basic axioms of what a CGM means. When we draw a graph, however complex, it is understood that the arrows in the graph capture *all* the conditional dependence relations. Drawing an arrow between two events is meaningful, and carries commitments between the two nodes. The Markov condition formalizes the idea that *not* drawing an arrow carries commitments as well – it implies particular relations of statistical independence. From a practical point of view, statistical independence can be expressed in terms of information gain. If the events A and B are independent, then information about A tells us nothing new about B. Mathematically, it means that the probability of A given any value of B is the same as the probability of A when B is unknown. So, the Markov condition tells us that information must be carried through inferences about direct parents.

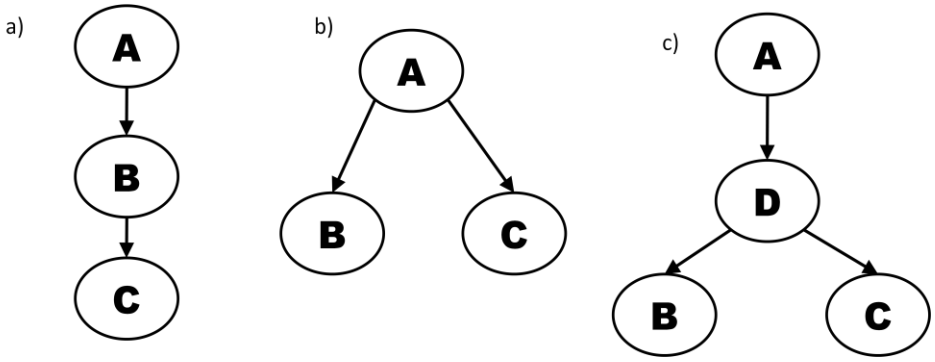

Figure 4: A chain structure (a) along with common cause structures following conjunctive fork (b) and interactive fork (c) patterns.

The Markov condition is most easily seen in a causal chain structure, where A causes B, which in turn causes C. See Figure 4a for an example. Knowing that A occurred

tells us something about whether C occurred. Also, knowing that C occurred tells us something about whether A occurred. But given B, information about C tells us nothing new about A, and vice versa. In Reichenbach's (1956) terminology, we say that B *screens off* A from C.

The Markov condition seems simple enough, but causality and inference are different things. For instance, an effect raises the probability of its cause. This is not because we have reversed the direction of causation or violated the Markov condition, but because the effect carries information about the effect that we can use in inference. By applying the Markov condition to a graph, we can tell which nodes give us information about other nodes, given the information we already have. Through these inferences, the Markov condition sometimes applies not just to parents, but to siblings: nodes that share a common parent. Figure 4b shows a common cause structure. In this case, A directly causes both B and C, each with some probability. Let's say we begin with no information, and then find out that C occurred. This raises the probability of B, because we can infer that A probably occurred as well, and A causes B. But if we already knew that A occurred, then C does not tell us anything new about B. This is because we already had all the information that matters, according to the Markov condition: We had information about C's direct parents. In this way, A screens off B from C, in a common cause structure like this one. This is known as *conditional independence*: B and C are dependent on each other given no information, but conditioned on A, they become independent. Note that this particular kind of conditional independence only

occurs when A is the direct parent of B and C. An example is shown in Figure 4c: D is an intermediate event that connects both B and C to A. If we know A, but not D then B and C can still give us information about each other, because they tell us about D. Salmon (1984) calls this an *interactive fork* as distinguished from the *conjunctive fork* of the direct common cause. Note that there must exist some indeterminism in the relation AD for the two to types of forks to be different in any statistically relevant sense. The deterministic limit of both cases Salmon calls a *perfect fork:* Both effects always occur given the common cause.

Overall, the Markov condition permits us do inference from nodes other than direct parents, but only *through* inferences on the parents. When direct parents are known, they are all that matters. It is important to emphasize that the Markov condition is not optional to CGMs. It is an integral part of what a CGM means. CGMs carry information about conditional dependence, information that only has meaning in the background context of independence that is carried by the Markov condition. Drawing a CGM without the Markov condition is like drawing with black ink on black paper – not much knowledge is gained.

Interventions

An important and valuable property of CGMs is their ability to formally represent the results of exogenous interventions. By 'exogenous' we do not necessarily mean an action that is not caused, but at least an action that is not caused by events in the graph. The mathematical implication of exogenaeity is that the event in question is

independent of other events. Without a formal mechanism for making sense of such actions, interventions are problematic. A canonical example (e.g. Gopnik et al, 2004) is that smoking causes both cancer, and yellow teeth. This means that both yellow teeth and smoking are correlated with cancer. Using CGMs, we can represent this using a simple common cause structure like Figure 5. Using the Markov condition, we can read from the graph that reducing smoking will reduce the probability (and, in a large population, the incidence) of cancer. But intuitively, teeth-whitening campaigns are unlikely to meet with much success in preventing cancer. This happens even though, in the absence of knowledge about smoking, yellow teeth can rationally help us infer the probability of cancer. There is an obvious asymmetry here. Properly formalizing the idea of intervention will help us make sense of the asymmetry.

Pearl (2000) provides a solution. We can formalize an intervention through what he calls 'graph surgery.' This is shown in Figure 5 as the 'Do' operation. If we wish to intervene on a node 'A', we first sever all connections between A and A's direct parents. (Note that because of the Markov condition, this implicitly severs connections among A's siblings, and the other parents of A's children, as well, if such indirect relations go through A.) Second, we set the value of A according to the intervention. We then use this modified graph to do inferences about the probability of the effects.

We can apply this to the smoking example, and see that it produces a sensible outcome. For instance, if we intervene on 'smoking,' setting it to 'no,' we do not need to do much work, because smoking does not have parents in this graph. We can then read

off the graph that cancer rates will be reduced. But if we intervene on 'yellow teeth,' we must sever the connection between yellow teeth and smoking. In this modified graph, the presence of yellow teeth tells us nothing about cancer, because we have severed the connection (through the common parent) between yellow teeth and cancer.

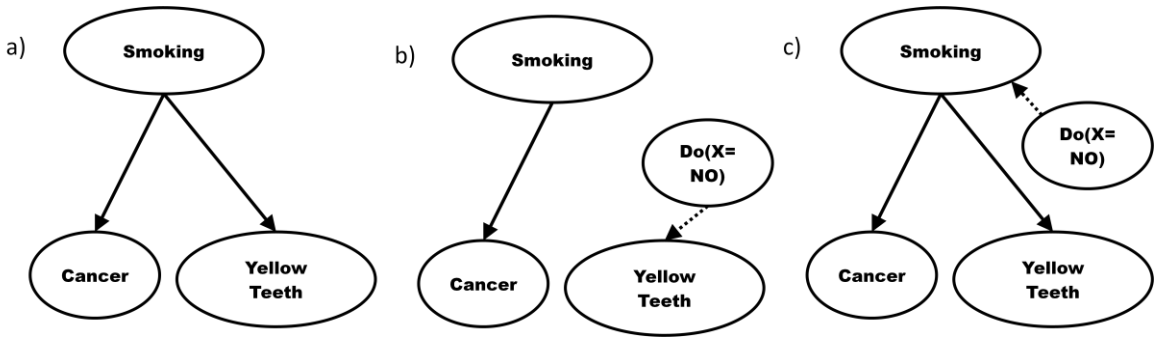

Figure 5: Examples of interventions on CGMs.

Pearl's (2000) complete 'calculus of intervention' has some additional complexity, including the ability to intervene on groups of nodes, and the ability to represent counterfactuals. This chapter will not discuss the finer points of these issues, on which some debate still continues. The basic notions presented here, will suffice to interpret some experimental data later on.

A problem: The hypothesis space of possible graphs

So far, this chapter has outlined a set of formal mechanisms, that together constitute CGMs: nodes, edges, functional forms (a small sample of which we have outlined), the Markov condition, and the idea of an intervention. The resulting formal system is extremely powerful. In fact, as we will see in this section, it is *too* powerful to be practically used, and needs additional constraints to rein it in. Unlike in the case of

the Markov condition, there is room for alternative approaches here – exactly how we rein in CGMs is a matter for debate and experimentation.

Too see why unrestricted CGMs are too powerful, consider the car example again. Turning the key caused two things to happen: the lights dimmed (most of the time) , and the car started (some of the time). At first we used the relatively simple graph shown in Figure 1 (and shown again in Figure 6) to represent this. Figure 6 also shows two other graphs that we could have used to represent the relation. These graphs are permitted because there is nothing to stop us from introducing arbitrary amounts of intermediate structure between a given cause and effect. The simple graph shown in 6a is permitted of course, but we could use Figure 6b, which seems rather sensible, or we could use the monstrosity shown in Figure 6c. Intuitively, Figure 6c seems wrong, but so far we have no formal principle that allows us to rule it out, because it captures the conditional probabilities just like the other graphs.

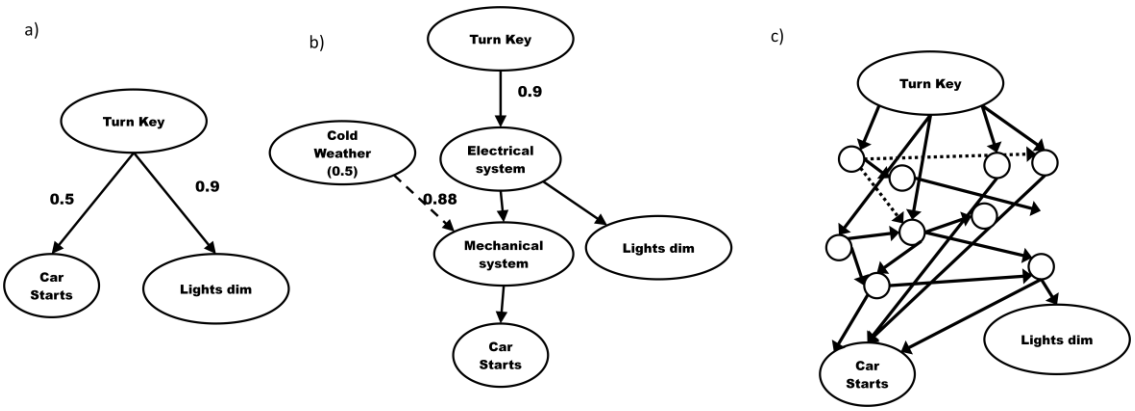

Figure 6: Three different graphs from the hypothesis space of graphs that capture the car example. Note that because of space constraints, not all functional forms are shown for (c).

One way of thinking about this problem is using the idea of a *hypothesis space*. A simpler example will illustrate: You are giving out candy on Halloween night, and you

need to decide how much candy to have available. You want there to be only a small probability (say 1 in 100) that you will run out of candy. For simplicity, let's say you give one piece of candy to each child. On past Halloweens, an average of twenty children have come to the door. But it is also theoretically possible that any given number of children will come to the door. For instance, there is some small possibility that 347 children will come to the door this evening. But this possibility is so remote that you would never buy 347 pieces of candy. One approach to this problem is to say that there is a hypothesis space of possibilities, each hypothesis representing a number of children that could come from the door. We can use a function (a Poisson distribution is often used in these cases) to assign a probability to each member of the hypothesis space. The Poisson distribution assigns a high value to the possibility that 20 children will come to the door, but a nonzero value to every other value, even 347. Even over the infinite space of possibilities, these probabilities sum to one, because the probability of very high numbers becomes vanishingly small. Intuitively, we can say that unlikely hypotheses are given fewer votes in our decision. To solve our problem, we just need to calculate the amount of candy that we need to make sure that only 1/100th of the probability won't be covered. One approach is to work through the hypothesis space, from zero up to the point where we have collected 99 per cent of the probability 'mass' (or, if you like, 99 per cent of the votes). Incidentally, the solution to this problem using a Poisson distribution is about 31 pieces of candy. [2] There are simpler approaches to this

[2]Because $\sum_{x=0}^{30} \frac{20^x}{x!} e^{-20} = 0.9865$, but $\sum_{x=0}^{31} \frac{20^x}{x!} e^{-20} = 0.9919$. In MATLAB:

specific example, but this approach illustrates that infinite hypothesis spaces can be tractable.

To return to the problem of causation: For any given causal relation (such as the car) there exists an infinite hypothesis space of possible CGMs that we could use to represent the causal relation. Figure 6 shows only a few illustrative examples. To add to the problem, large sets of these graphs are statistically indistinguishable from each other. That is, they predict the same net functional relations between the main causes and final effects. For instance, we could always replace any relation A->B, with the relation A->X->B, as long as the net effect was such that A and B obeyed the same statistical relations with each other. We could recursively apply this procedure to create whole menageries of statistically identical graphs. On the other hand, some complex graphs actually do make different statistical predictions from each other, even though they capture the same core relations between main causes and effects. For instance, Figures 6a and 6b both say 'turning the key cases the car to start' and 'turning the key causes the lights to dim.' Only Figure 6a, however, allows for the possibility that the car could start, without the lights dimming. This is because there is a deterministic relation between the node 'electrical system' and 'lights dim,' but 'electrical system' is a necessary connection between 'turn key' and 'car starts.' In Figure 6a however, it is

---

poissinv(0.99,20)=31

entirely possible to have the car start, without hearing the sound of the engine turning over. For instance, the left edge could fail, while the right edge succeeds.

Note that the preceding analysis actually understates the problem of the enormous hypothesis space created by CGMs. The analysis uses a restricted set of functional forms (the Noisy-OR and inhibitory forms described in the earlier section) to make the problems easier to illustrate graphically. This is just for convenience; as we have described them now, CGMs give us no formal reason to restrict the functional forms like this. Also, note that showing a few examples cannot convey the sheer number, and complexity, of the wildest graphical models possible under CGMs. Most individual members of the hypothesis space would not fit on one page.

We cannot proceed until we solve the problem of the intractable hypothesis space. For instance, we could try to do inference in the car example, trying to predict the probability that the car would start given that the lights would dim. Different graphs give different answers to this question, and we have no principle to tell us which subset to focus on.

Recall that in the Halloween example, we dealt with an infinite hypothesis space by using a distribution that assigned a probability to every member of the infinite set. The hypothesis space became tractable because we assigned vanishingly small (but nonzero) probabilities to unlikely possibilities. We then weighted the role that each hypothesis played in our decision, by its assigned probability. Another solution might have been to truncate the hypothesis space, for instance by considering only the most

likely hypothesis (20 children, the mean of previous years.) Note, of course, that this approximation would underestimate the amount of candy (31 pieces) that the Poisson analysis thought was appropriate. Below, we will see that most applications of CGMs currently use a strategy analogous to truncating. Chapter 2 proposes a strategy that is more like assigning a probability to every member of the space.

One solution: Minimality

Most existing approaches to causal inference (e.g. Pearl, 2000/ 2009; Spirtes, Glymour, & Scheines,1993/2000) solve the problem of the intractable hypothesis space by using a *minimality* constraint. Though different theorists phrase it in slightly different ways, it amounts to the same thing: Among graphs that capture all the dependencies in the data, consider only those that have the fewest possible edges. It is possible that there will be multiple minimal graphs, but it is at least guaranteed that there will not be infinitely many of them. From the point of view of the hypothesis space described above, minimality is equivalent to assigning probability zero to any graph that has anything greater than the minimal number of edges, and dividing the probability among the remaining, minimal models. The scheme for dividing this remaining probability differs between algorithms, but these differences are irrelevant to the arguments made in this thesis. Often there is only one minimal graph.

In principle, minimality is uncontroversial – it is in practice that it presents problems. When certain assumptions are met, minimality amounts to saying that we should avoid adding a specific kind of complexity: the kind that is both completely

unnecessary, and makes no difference. The assumptions required for this lack of controversy are strong, however: First, that the 'data' the graph fits is the long term limit of the statistical distribution of events; we actually know whether any two events are correlated or independent in the long run distribution. That is, we presume that we have complete knowledge of even the subtlest violation of the Markov condition. Second, minimality sometimes assumes what Spirtes et al. (1993/2000) call 'sufficiency' – that we have data on all the events that make up the graph. With these two assumptions in place, minimality can't go wrong. In the real world, however, and especially in human reasoning and learning, it is rare that either of these two assumptions holds. We usually operate with limited data. In fact, we sometimes operate with NO data – we will deal with cases below in which subjects are just told 'A causes B' and asked to make probability judgments. We also usually do not know about the status, let alone the existence, of hidden causes and intermediate events. In fact, the existence of hidden causes is often the interesting question we are trying to answer – assuming we already know all the causes, won't help. In practice, researchers often try to apply minimality as a simplifying assumption, to cases where these assumptions do not hold. This is when things go wrong.

Minimality is one way of expressing Ockham's famous principle of simplicity in scientific hypotheses. (As we will see in Chapter 3, there may be alternative ways of measuring simplicity.) One virtue of a simplicity prior is philosophical: many of us share the intuition that simpler hypotheses are more likely to be correct. Another is formal:

For any given level of complexity (in this case, defined by the number of edges) there are fewer hypotheses to consider at lower levels of complexity. From minimality's point of view, if we consider graphs with five edges, there might be dozens, but if we consider graphs with only three, there might be only one. This means there is less ambiguity in the causal relations, and lets us set about the already difficult task of deducing the functional form of the causal relations, with some degree of assurance. Related to this property is the computational convenience of minimality: It is difficult enough to construct search and inference algorithms that recover minimal structures. Allowing non-minimal structures quickly gets out of hand, presenting new challenges. In fact, even given strong arguments against minimality, the strength and efficiency of its search algorithms may still recommend minimality in many situations, for instance in machine learning over large data sets where the direction of causation is not known. In psychology, computational convenience is less relevant: If the data show that people are not using what we consider to be the most tractable algorithm, they must be using some other method, even if we do not yet understand what it is.

Minimality in Action

Several kinds of algorithms recover minimal structures. These include the SGS and TETRAD classes of algorithms from Spirtes et. al. (1993/2000), and the IC and IC* algorithms from Pearl (2000/2009). These are known as *constraint-based* algorithms because they use a set of rules to converge on a single minimal model at a time. Other models that obey minimality include some Bayesian approaches (e.g. Heckerman, Meek,

& Cooper, 1997) that explicitly consider the larger hypothesis space of possible models, but only those that are as simple as possible. (Note that there are some Bayesian models that go beyond minimality – they are discussed below.) Though they differ in the details of implementation, and sometimes converge on different minimal models, all the models just mentioned share a key principle: Introduce new edges, and new nodes, only when it is necessary to do so in order to preserve the Markov Condition. This conservative mechanism commits these models to a common set of predictions.

The car example will serve to illustrate minimality in action. For the purposes of this section we will assume that the full model shown in Figure 3 (and shown again in Figure 7) is the 'correct' model; the data we observe is generated from it. You initially observe three events: turning the key, the car starting, and the lights dimming. Because you intervened on turning the key, you know that it must be a cause of the other events. You also know from interventions that dimming the lights is not an intermediate cause in a chain, on the way to the car starting. For instance, if you turn off the lights, turning the key will still cause the car to start, even though the lights do not dim.

Under a model that employs minimality, we draw all and only those edges that are necessary to prevent violations of the Markov condition. See Figure 8a for the result. If an edge did not exist between 'turn key' and 'car starts,' there would be a correlation between the two events that was not explained by the graph. We call such an unexplained correlation a 'Markov violation.' A similar argument applies for the edge between 'turn key' and 'lights dim.' Once these edges are in place, of course, we can

explain the correlation between ‘car starts’ and ‘lights dim’ – they have a common cause, so no edge is needed between them.

Once the graph is constructed, we assign functional forms to each causal relation. In this case, we assign generative, probabilistic forms to each relation, as shown in the graph.

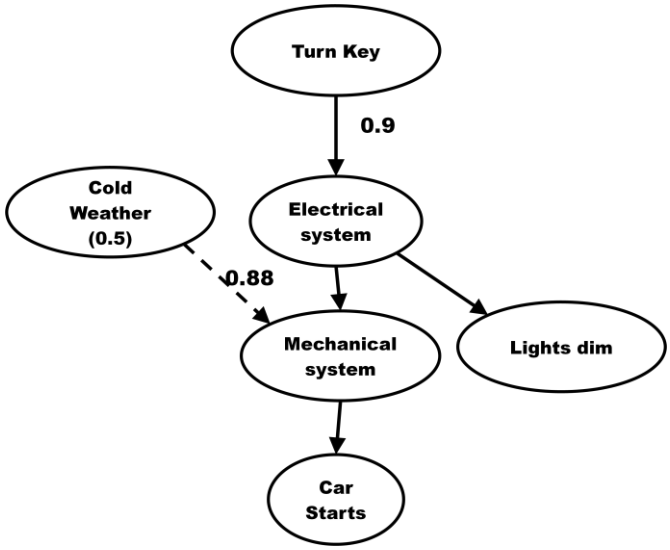

Figure 7: The model we will assume is the ‘correct’ model of the car example.

Note that we have not assigned any intermediate structure, as exists in Figure 7, which we have assumed for simplicity is the ‘correct model.’ This is because such additional structure is not needed to explain any Markov violation that exists in the data. Of course, this is not entirely fair to minimality, because the algorithm was given limited data, and was not told about the fact that an additional hidden variable (‘cold weather’) was present. What happens if we remedy this?

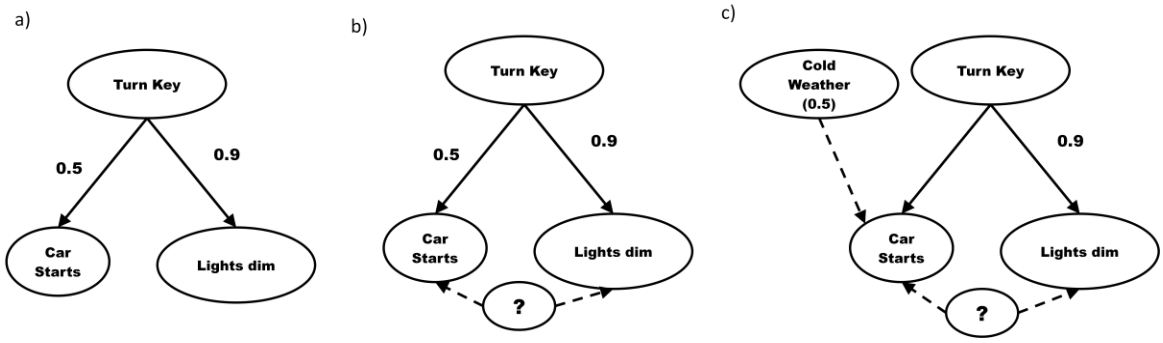

Figure 8: A series of graphs, showing what a constraint-based model would infer given progressively more data.

We can first remedy the problem of missing data. Over time, as more data accumulates, a new Markov violation will start to appear: A correlation will exist that is not captured by Figure 8a. We will find an additional, more subtle correlation between 'lights dim' and 'car starts' that is not captured by the fact that they share the common cause 'turn key.' This is because in the 'true' model, there is an interactive fork: it is not possible for the car to start unless the electrical system is working. A failure in the electrical system will disable both the starting of the car, and the dimming of the lights, creating a correlation, even when we control for turning the key.

Over time, the occasional failure of the electrical system will create a detectable Markov violation: 'lights dim' and 'car starts' will not be independent, even given 'turn key,' even though under the simplest model in Figure 8a, they should be independent. This licenses us, under minimality, to remedy the situation through the introduction of new nodes and edges. Different models handle this in slightly different ways, but all amount to making the smallest possible adjustment to the model, to accommodate the Markov violation. Figure 8b shows how the TETRAD algorithm handles the situation: By

introducing a hidden common cause of the two effects. Now, through the use of more complex functional forms, we can explain the correlation. Note that it can take a great deal of evidence for such correlations to reach statistical significance, especially for rare events like the failure of a car's electrical system. When dealing with real world data, a minimal algorithm may require a large amount of data before overturning an overly simple graph.

The second piece of injustice we visited on the minimal model, was that we did not tell it about the existence of the hidden variable 'cold weather.' If we explicitly introduce 'cold weather' as a variable, the algorithm will immediately start to notice a correlation between 'cold weather' and 'car starts.' A minimal model will be justified in introducing an edge to explain this correlation, as shown in Figure 8c. We introduce an inhibitory relation between 'turn key' and 'car starts.' This is another way that minimal models create more complex graphs: When we introduce new variables, minimal models can appropriately incorporate them into the graph.

What happens if we never tell the minimal model about 'cold weather'? In this case, it will never infer the existence of a hidden disabler. This is because it will be perfectly content to keep the probabilistic edge connecting 'turn key' and 'car starts' – after all, it does explain the correlation. Markov violations are the irritants that cause minimal models to grow. Positing a probabilistic causal relation is both simple and sufficient to remove the irritant, so there is no motivation for a minimal model to posit a more complex, deterministic relation.

At this stage, we have constructed minimality's answer to the question posed in the introduction (namely, how we develop more complex representations of causal structure). As a model of human causal learning, minimality holds that people should infer intermediate structure (mechanisms) gradually, over time, as statistical evidence accumulates. With knowledge of few events and little data, they should begin with simple models with few edges. As knowledge increases and data accumulates, they should add edges and nodes only as necessary to explain correlations. The Markov assumption means that independence is the default assumption in CGMs. Information about dependencies is carried by edges; by minimizing edges, minimality commits us to a default expectation of independence. Furthermore, minimality carries a strong motivation to expect simple probabilistic causal relations, rather than more complex deterministic ones. Hidden causes are incorporated in the models only as they are revealed – they are not expected in advance. In summary, minimality prescribes that people should be surprised to find both interactive forks and hidden causes. Below, we will see that evidence that in some important situations, people are not surprised – in fact they seem to expect these structures.

Problems with Minimality

As a model of human causal reasoning, minimality has three major problems. This section will describe them in turn, in increasing order of strength. We begin with a quick summary: The first, and probably weakest, criticism, is that minimal models seem to be theoretically incompatible with the idea of a representation of causal mechanism.

Mechanism is minimized away. Ultimately, however, minimal models do present a (somewhat deflationary) account of mechanism, at least in the sense of how we infer hidden causes and intermediate structure, so the tension can ultimately be resolved through empirical evidence. A second, stronger argument concerns such evidence: It seems that even young children are determinists, in the sense that they prefer complex deterministic structures to simple probabilistic ones. Evidence for this includes the fact that children infer the intervention of a hidden confederate given a variable causal relation, but not given a stable one. This appears to violate minimality. The third argument is the strongest: Human reasoners show a robust *nonindepence* effect when doing causal reasoning with little data: They are liberal with the correlations they posit, while minimality predicts they should be conservative.

Mechanism

The question we are concerned with is causal mechanisms: How do people learn about the events that connect causes and effects? A proposed model must their account for mechanism-based reasoning, or else explain away the evidence that people do reason in mechanism-based ways. The most obvious and hopefully uncontroversial piece of such evidence is that people do have knowledge of intermediate events. Empirically, for instance, Rosenblit and Keil (2002) show that when asked how an artifact (for instance, a helicopter) works, adults and children can generate descriptions of intermediate events. While these representations vary in their depth and detail, they do exist. Furthermore, there is evidence from the ways in which people engage in causal

attribution: determining whether an event actually causes another. Ahn, Kalish, Medin, and Gelman (1995) show that when asked to engage in causal attribution, adults are more likely to ask about causal mechanisms, than about statistical properties. Also, Shultz (1982) shows that children reason about mechanisms in causal attribution: They recognize that the cause of an event is the one that could actually transmit force or energy from cause to effect.

If we take minimality seriously as a principle of human causal learning, then it predicts that human representations of causal structure should be highly resistant to positing mechanisms, especially when encountering a novel causal system. All such structure must be motivated through painstaking statistical deduction. While it is possible that such deductions are the operations through which we arrive at representations of mechanism, this seems highly implausible.

Nonetheless, minimality does provide an account of the learning we are interested in. Even if we grant the plausibility of the proposed operations, minimality still makes specific predictions about the *way* in which such learning should progress. Namely, people should incorporate hidden inhibitors when they observe them, but not expect them in advance, and should accept interactive forks only after a great deal of statistical evidence. In the next two sections, we will compare these predictions to empirical evidence.

Determinism

Recall that we saw that minimality is conservative about positing hidden structure to account for specific functional forms. For instance, in the car example, given that turning the key causes the car to start 50 per cent of the time, minimality will happily infer a probabilistic causal relation with strength 0.5 between the two events. It strictly prefers this representation to the possibility that some unobserved event (such as cold weather) is inhibiting the causal relation 50 per cent of the time. Minimality strongly prefers simplicity, to the point that it will accept a variety of functional forms to achieve it. There is evidence, however, that even children have a converse preference: They prefer deterministic causal structures, to the point that they will sacrifice simplicity to maintain them.

There are several things that are meant by 'determinism' in the literature, so it is important to clarify what sense is meant here. Some researchers (e.g. Bullock, Gelman, & Baillargeon, 1982) take determinism to mean the idea that every event has a cause. While preschoolers seem to believe this idea (Bullock et al., 1982), it is a weaker principle than is needed to infer intermediate structure. That is, a probabilistic causal relation is sufficient to explain an event under this formulation of determinism.

Another form of determinism is probably too strong: The idea that causes always produce their effects. Not only is this manifestly wrong to anyone with experience operating in the world (we can do interventions that have different results on different

trials), but experimental evidence indicates that even infants will continue to intervene on events that produce their effects only sometimes (e.g. Watson, 1979).

A third form of determinism is what is meant in this thesis: The idea that *direct* causes always produce their effects. This means that variable results indicate that some intermediate structure exists mediating between cause and effect. This is sometimes known as *Laplacian* determinism, after Laplace, (1814/1951), who claimed that while the world sometimes appears random, this is due only to our ignorance of the true, deterministic causes. Under this view, causation itself is deterministic, in the sense that it always produces the same results; apparent variability comes from a lack of knowledge of the true causal structure and the hidden events.

Note that this section does not endorse Laplacian determinism as a metaphysical claim. In fact, recent findings in quantum physics suggest that the universe may have randomness built into its most fundamental structure. But, as Schulz and Sommerville (2006) point out, "a belief in causal determinism need not be metaphysically accurate to be functionally adaptive" (p. 429). They argue that determinism is adaptive because it leads us to look for hidden structure in causal relations. This thesis will use the term 'determinism' to refer to the hypothesis that humans have a default assumption of Laplacian determinism (direct causes always do the same thing, variability is because of intermediate structure), whether or not they actually should.

We can read this definition of determinism back in to the language of CGMs, in order to help it generate specific predictions. Laplacian determinism amounts to a

preference for functional forms in which the probability of the effect given the cause is 1. This preference should persist even when we must posit a non-minimal causal structure in order to accommodate it.

Evidence suggests that humans are determinists in this sense. For instance, Goldvarg and Johnson-Laird (2001) show evidence that when told about a causal relation between A and B, participants initially expect that the situation in which A occurs, but B does not, is impossible. When told that A occurred but B did not, people readily generate an explanation when prompted to explain the failure (Walsh & Sloman, 2008). Legare, Gelman, and Wellman, (2010) show that preschoolers readily generate such explanations as well. Such explanations rarely involve probability – instead they usually involve events or properties that disable the causal relation. This evidence from explanations is weakened by the fact that the explanation is usually elicited – it does not show that people fail completely to represent causal relations probabilistically, only that they *can* represent them deterministically. Further evidence shows that people prefer deterministic causal relations: Lu, Yuille, Liljeholm, Cheng, and Holyoak, (2008), show that a prior distribution that favors deterministic causal relations fits human data better than a prior distribution that has weaker preferences about functional forms. That is, humans endorse a causal relation significantly more strongly when the cause appears to always produce the effect; models that do not take determinism into account do not fit the strength of this preference as accurately. Overall, there is relatively weak evidence for determinism in adults; more is needed. The strongest evidence comes from an

experiment on 4- and 5-year-olds, conducted by Schulz and Sommerville (2006). This section will focus on Schulz and Sommerville's experiment, because it is most relevant to the overall arguments about minimality.

Schulz and Sommerville showed children a machine that lights up and plays music when a switch is pressed. The machine also had a ring on the top surface. Children saw three trials in which pressing the switch caused the machine to activate. They then saw three trails in which the ring was removed; on these trials, the machine did not activate when the switch was pressed. This gave strong evidence that removing the ring was an inhibitor. At this point, the experimenter and confederate changed places, and the confederate pressed the switch eight times. This resulted in a sporadic pattern: the machine only activated in two, nonconsecutive trials. The experimenter then revealed that she had been hiding a flashlight in her hand the whole time. She placed the flashlight next to the machine (the ring remained on the machine), and announced that she was about to press the switch. Children were asked to inhibit the activation of the machine. ("Can you make it so that the switch won't work and the toy won't turn on?")

Note that at this stage children have two choices: They can choose to remove the ring, which they have previously seen inhibit the causal relation. This is the reliable, conservative action. They can also choose to intervene on the flashlight. Intervening on the flashlight indicates that children inferred that the flashlight was inhibiting the

machine during the sporadic period; the ring was visible and present on top of the machine.

Results indicate that most children chose the flashlight. This was in contrast to a condition in which the machine activated eight out of eight times for the confederate; children in this condition tended to choose the ring. They also chose the ring when the sporadic activation was explained away by showing that the confederate was not pressing down hard enough on the switch.

We can explain the data as follows: When children observe the sporadic activation of the machine, they accommodate the variability by positing a hidden inhibitory cause. When the flashlight is revealed, it fits the posited pattern of activation, and so children accept it as an inhibitory cause. They indicate this acceptance by choosing the flashlight as an inhibitor. Their acceptance is so strong that they do so even though they have never actually seen the flashlight inhibit the causal relation, and they have an alternative choice that has been shown to be effective in inhibiting the relation.

What does Schulz and Sommerville's evidence mean for minimality? In order to explain children's behavior in this experiment, we require an account in which children posit a hidden inhibitor, even when they have not seen any trials on which the supposed inhibitor was present. This violates minimality. According to minimality, children should use a complex functional form, not a complex causal structure, to accommodate the variability. If children were using complex functional form, they should choose the ring on all three conditions. In the sporadic condition, the sporadic activation should just be

indicative of a probabilistic relation – information that is irrelevant to the existence of hidden inhibitors.

This is the most direct evidence that exists in the literature on children's determinism. While this evidence is strong, replications and extensions would significantly improve determinism-based arguments. Chapters 5 and 6 will show evidence that develops, and further supports these conclusions in preschool-aged children. For now, the available evidence suggests that people prefer to posit intermediate, deterministic structure, than accept a probabilistic relation. This is in violation of minimality. However, because the evidence is not particularly extensive, more evidence is needed for determinism.

Nonindependence

The strongest argument against minimality is the phenomenon of *nonindependence*. This refers to human subjects' tendency to posit correlations among collateral effects of a common cause, when minimality says that they should not. The strength of the nonindependence argument comes from the fact that nonindependence effects have been replicated by several researchers using a variety of scenarios and methods, together with the fact that minimality's prescriptions are both so clear and so counter to human data.

An example of nonindependence comes from Walsh and Sloman (2008): Subjects were told that jogging causes each of weight loss, and increased fitness. Subjects received no data on strength or co-occurrence, just the endorsement of these

two causal relations. The minimal structure given such a cover story is shown in Figure 9: this is the familiar minimal common cause structure. According to minimality, no more complex structure can be licensed, because that would require enough data to detect a Markov violation, and subjects were given no data. The minimal graph should be their default assumption.

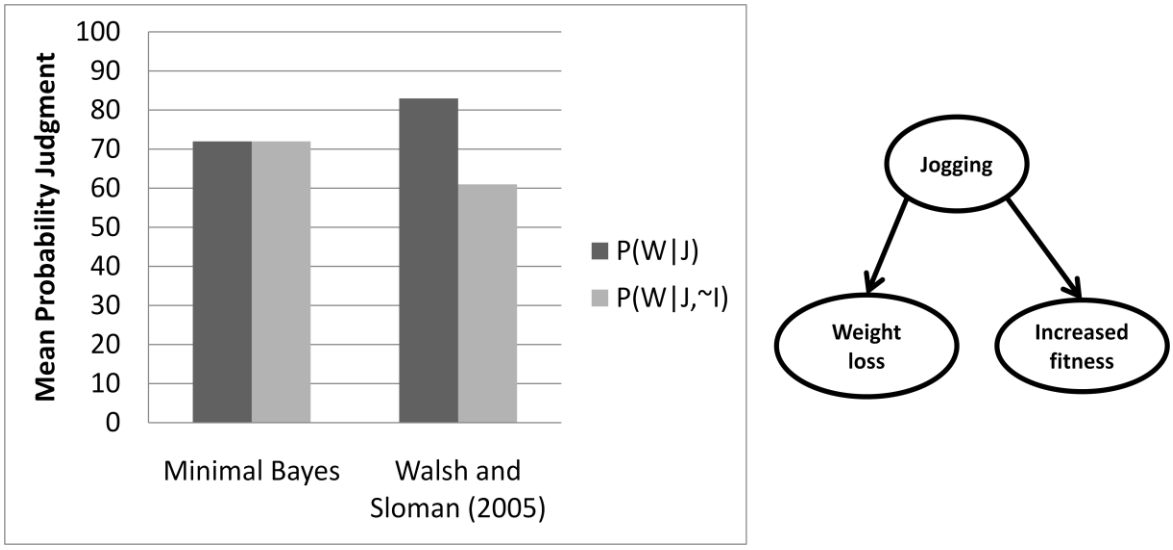

Figure 9: Walsh and Sloman's participant's judgments about jogging, fitness, and weight loss, along with minimal predictions

In one experiment, subjects were told about a specific person, Bob, who started jogging. Subjects were asked to judge the probability that Bob's fitness had increased. The average judgment for these types of questions was around 0.85 (shown as the bar P(W|I)[3] in Figure 9.). Subjects were then told that Bob had NOT lost weight, and then asked the question about fitness again. Subjects tended to lower their probability judgments, to about 0.63 (shown as the bar P(W|J, ~I)). Information about one

[3] This means: The probability of W (weight loss) given I (increased fitness). Such standard notation will be used throughout the thesis.

collateral effect caused them to change their judgment about another. Given a minimal structure like that shown in Figure 9, this violates the Markov condition: Because the cause is given, the two effects should be independent, meaning that information about one should tell us nothing about the other. Figure 9 shows minimality's predictions, next to human data. Minimality cannot account for the significant drop in probability judgments.

Walsh and Sloman replicated this basic effect across a series of experiments. Other researchers have replicated this effect as well, often in ways that are even more problematic for minimality. For instance, Rehder and Burnett showed that nonindependence holds in the context of categorization, when one central feature of a category caused three other features. In this case, subjects were specifically told the functional form of the causal relation: The central feature caused each other feature with strength 0.75. Subjects still showed a nonindependence effect, as shown in Figure 10. This 'staircase' pattern, in which judgments progressively increase as the presence of collateral effects is added, is the hallmark of a nonindependence effect. Again, Rehder and Burnett replicated this effect across a series of cover stories. They even showed nonindependence effects when using blank predicates (literally, "A causes B") and when the independence of the mechanisms was explicitly emphasized in the cover story. These data are also shown in Figure 10.

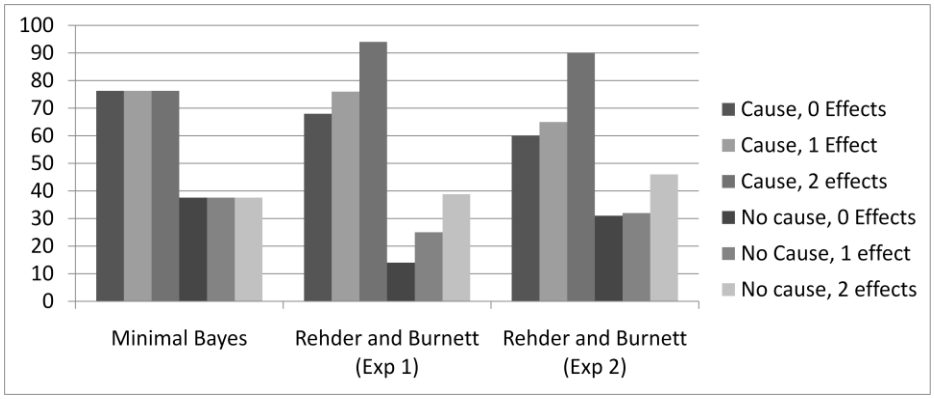

Figure 10: Data from Rehder and Burnett. Experiment 2 used blank predicates.

The robustness of these effects lets us address a number of possible defenses of minimality as a psychological theory. For instance, Walsh and Sloman's experiments, it might be argued, are not directly applicable to minimality because they asked subjects to make probability judgments just from cover stories, not from data. But Rehder and Burnett found the same effect when subjects were given information about the strength of causal relations, and the rates of the causes and effects. Another possible objection is that subjects might have specific mechanism knowledge, arrived at through proper minimal statistical deduction, that led them to expect correlations in cases like jogging. However, Rehder and Burnett replicated the effect using novel and even blank predicates, where such deductions could not have occurred. Even explicitly emphasizing independence did not seem to work. Even given minimal information, people's default assumptions about common cause scenarios, do not conform to minimality's predictions.

Solutions to Nonindpendence

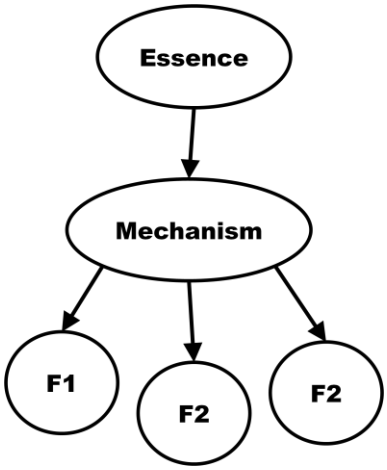

Figure 11:Rehder and Burnett's (2005) 'Underlying Mechanism Model'

Apparent violations of minimality raise the question of what principle people *are* using. One solution is to propose that people are explicitly representing intermediate structure in common cause scenarios. An example is the 'underlying mechanism model' (UMM) proposed by Rehder and Burnett (2005). They propose that people posit a hidden cause between the essence and the features (the mechanism) by default when representing a category with multiple features. While this model qualitatively solves the problems presented by Rehder and Burnett's experiments, attempting to generalize this insight opens a Pandora's box. Recall that without minimality, we had no way of constraining the hypothesis space of possible causal models. To the extent that the underlying mechanism model generalizes, it abandons minimality, leaving us without a constraining principle. Furthermore, the UMM does not tell us exactly how to set up the functional forms within the more complex structures it posits. For instance, how much of the variability should we attribute to the mechanism, versus the relation between the

mechanism and the individual features? This means that it has difficulty making specific quantitative fits and predictions. Despite these shortcomings, Rehder and Burnett's idea of introducing intermediate structure shows promise for developing an alternative to minimality. The model proposed in Chapter 2 will solve these shortcomings, and recover the UMM along the way.

Another solution to minimality's conflict with nonindependence, is to make a small amendment to the minimal structure. For instance, we could say that given any common cause structure, there exists by default a hidden source of common inhibitory noise. This hidden cause explains the correlation, while preserving the Markov condition. This approach is taken by Mayrhofer, Hagmayer, and Waldmann (2010). Fitting the strength of inhibitory noise allows them to fit human nonindependence data well. This approach is least problematic, but also (as the next section will argue) lacks explanatory power.

An interesting experiment

Mayrhofer et al (2010) conducted an experiment that shows some of the most interesting aspects yet of nonindependence: The degree of nonindependence depends in subtle ways on the way the mechanism is described. Mayrhofer et al manipulated the cover story slightly between conditions, changing only the description of the mechanism. They found that this changed the degree of nonindpendence they observed. This has important implications for the ways in which we might model nonindependence.

Mayrhofer et al told subjects four telepathic aliens, one we will call the 'cause' alien, and three we will call the 'effect' aliens. According to the cover story, when the cause alien thought of food, he could cause the three effect aliens to think of food. Mayrhofer et al slightly manipulated the cover story between participants: In the 'sending' condition, participants were told that the cause alien sent his thoughts to the effect aliens, but sometimes had trouble concentrating. In the 'reading' condition, participants were told that the three effect aliens read the thoughts of the cause alien, but each sometimes had trouble concentrating. In an attempt to replicate nonindependence effects, they also manipulated within subjects the number of other effect aliens were thinking of food, given that the cause alien was thinking of food. Participants were asked to judge the probability that one effect alien was thinking of food, given that the cause alien, and 0,1, or 2 other effect aliens were thinking of food.

Results are shown in Figure 12. Note the familiar staircase pattern, indicating that Mayrhofer et al replicated the nonindependence effect found in previous experiments. Statistical tests indicate a significant effect of the number of active collateral effects, in violation of minimality. Note also that the nonindependence effect was significantly stronger in the 'sending' than in the 'reading' condition. The description of the mechanism changed the degree of nonindependence.

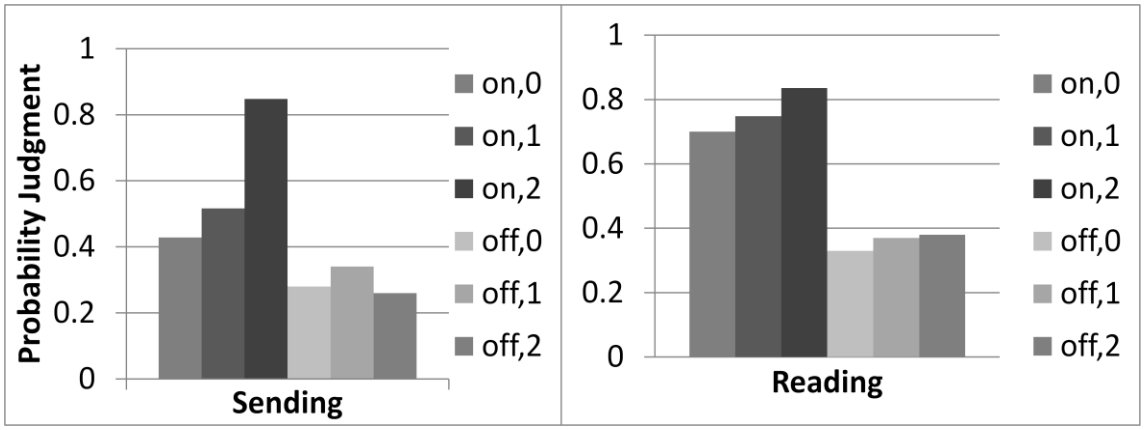

Figure 12:Data from Mayrhofer et al (2010). The legend indicates first whether the main cause was on or off, the number of collateral effects present. For instance, 'on,1' indicates that the cause alien was thinking of food, and only one other effect alien was.

Mayrhofer et al fit this effect by fitting a parameter: In the sending condition, the inhibitory noise was strong, but in the reading condition, the inhibitory noise was weak. While this account fits the data well, quantitatively speaking, it leaves unanswered the question of *why* the effect was stronger in one condition than in the other. The model could also have accounted for the converse effect just as easily – that is, if the nonindependence effect was weak in the sending condition and strong in the reading condition. Their account thus lacks explanatory power. This can be seen when we try to generalize their explanation to a novel example that does not involve transmission and reception, but does involve information about mechanism. We will look at empirical data for such a generalization, in Chapter 3.

Overall, nonindependence presents a serious problem for minimality. The Markov condition tells us that only edges carry dependence – all other relations are independent. Minimality, by minimizing the number of edges, predicts that people should tend to underestimate dependence. Instead, they seem to overestimate it.

Furthermore, solutions to nonindependence require much further development in order to be general or coherent. Developing such an alternative will be the focus of Chapter 2.

Prior knowledge and causal grammars

Some approaches allow for isolated exceptions to minimality. For instance, one move that is often made when discussing CGMs is that 'prior knowledge' (e.g. Glymour, 2001) can influence the kinds of causal structures we posit in a given situation. One relatively uncontroversial example of such knowledge is the constraint that causes must precede their effects in time. We need not use statistical techniques to deduce the direction of causation, if we know the true temporal order. It is easy to see how we could use such a move to deal with nonindependence and determinism effects: In some situations, we have prior knowledge that intermediate structure and hidden causes exist. The problem with such an approach is that both determinism and nonindependence effects have been replicated with non-experts on novel causal systems. For instance, recall that Schulz and Sommerville used a novel machine, and Rehder and Burnett showed nonindependence effect even with blank predicates. Thus, the 'prior knowledge' we expect would have to be at a very high level, and applied very generally. Such arguments weaken the minimal account, especially if minimality seems to be honored more in the breach than in the exception. They also fail to formalize how such prior knowledge is learned and how it is applied.

Some researchers have proposed frameworks in which the learning and application of such prior knowledge could be operationalized. One notable example is

the framework for hierarchical Bayesian models of causal structure proposed by Griffiths and Tenenbaum (2007), that uses a 'causal schema' for hidden structures. Under such a model, causal events are fit into a learned template that can include hidden intermediate structure. One example that Griffiths and Tenenbaum (2007) use is that of diseases. We might learn that smoking causes bronchitis, which causes coughing. We might also learn that stress causes heart disease, which causes chest pains. From these and other examples, we might extract a template: behaviors cause diseases, which in turn cause symptoms. When we hear of a new symptom-disease pair that obeys a conditional dependence relation (for instance, drinking dirty water causes fever) we would represent the causal relation only by positing a new, as yet undiscovered disease. Because the disease is unobserved, and not necessary to explain any Markov violation, it is technically a violation of minimality to posit a disease in this case. But because of the causal template we have learned, there is a formally coherent justification for this exception to minimality.

These models are known as 'hierarchical' because they propose that learning occurs simultaneously at multiple levels. At the same time that we learn that smoking causes coughing (the level of a *graph)*, we are also learning that behaviors cause diseases, which cause symptoms (the level of a *template* for graphs). We might potentially be learning even more general pieces of knowledge, like the fact that causes tend to precede their effects. Presumably such high-level facts have been learned by adulthood, if they are learned at all. One possibility opened up by the hierarchical

Bayesian framework is that early on in development, there is a great deal of development going on at the highest levels. Griffiths and Tenenbaum (2007) propose that we look at the highest level of knowledge as a *grammar* out of which causal representations are structured. The analogy to language is deliberate; there have already been successful models that show how a linguistic grammar can be induced (e.g. Perfors, Tenenbaum, & Regier, 2006). Similarly, in causation, computational models have been implemented (e.g. Goodman, Ullman, & Tenenbaum, 2009) that show that the hierarchical Bayesian approach can learn at multiple levels simultaneously.

Interpreted within this framework, minimality lives not at the level of a schema, but at the level of a grammar. It is one of the rules out of which causal representations, be they schemas or specific models, are made. This means that even given a causal template that provides an isolated exception, minimality still holds by default in general. For instance, given a disease with multiple symptoms, we would initially assume that all symptoms were caused independently, given the disease. It would take time and data to learn that diseases tend to cause different symptoms through different classes of mechanisms. This opens the door to nonindependence effects. As long as minimality is still a default part of the grammar, such hierarchical models will still be subject to some of the same criticisms outlined above. While template-based models are a step forward in the sense that they provide for learning of richer intermediate structure, they still show the potential to underestimate the degree to which people posit intermediate causal structure.

What is really needed here is a coherent, operationalized alternative to minimality, at the level of the causal grammar being implemented. Currently no such alternative exists in the literature. Chapter 2 will develop such an alternative, using a generative edge replacement rule. Edge replacement is sympathetic to the hierarchical Bayesian framework, and can be interpreted as part of this project. It goes further, however, than other models in addressing the three problems outlined in this chapter. It will overcome the problem of nonindependence by positing interactive forks by default in causal representations. It will overcome the problem of determinism by forcing representations to be build out of a restricted, deterministic function set. Finally, we will see that the resulting formal and theoretical implications shed new light on the question of mechanism, rather than minimizing it away.

CHAPTER 2: THE EDGE REPLACEMENT MODEL

Edge replacement is an alternative to minimality. Recall that CGMs have a problem: because there is an infinite hypothesis space, we need some way of deciding which graph to use. While minimality solves this problem by using the simplest graph consistent with the data, edge replacement instead takes a generative approach: The model specifies a set of rules out of which all CGMs have to be made. As a scientific hypothesis, edge replacement amounts to the assertion that human representations of causal structure are built in a way that is similar, at the computational level (Marr, 1982), to these rules. This chapter will first outline the rules, then show how they explain some data that is problematic for minimality, particularly nonindependence and determinism effects. This will be followed by discussion of some of the more theoretical implications of its success. Before beginning this chapter, it is necessary to set up some mathematical background for readers who are not familiar with generative models.

Generative models and sampling

Edge replacement solves the problem of the large, complex hypothesis space using a generative approach. This section will introduce generative models using a simple example. Readers who are already familiar with generative models can skip this section.

Imagine we play a game, in which we flip a coin repeatedly. The first player flips the coin. If it is heads, the first player wins. If it is tails, the second player flips the coin. If it comes up heads, the second player wins; otherwise, the second player passes the coin

back to the first player. The process continues until someone tosses heads. In theory, the process could go on forever – we could keep flipping tails – but in the long run, the probability of any sufficiently large number of tosses approaches zero.

What is the probability of winning the game? It is intuitive that the first player has a slight advantage – he or she has a 50 per cent chance of winning on the first toss. Mathematically, the first player wins if a series of coin tosses has the first heads on an odd-numbered toss. So, the total probability of the first player winning is:

$$P(First\ Player\ Wins) = \frac{1}{2^1} + \frac{1}{2^3} + \frac{1}{2^5} + \frac{1}{2^7} + \cdots = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^{2n+1} = \frac{2}{3}$$

This is known as an analytic solution, because we arrived at an exact answer through a formal mathematical analysis of the problem. We could also have taken a *generative sampling* approach, which works like this: simulate the act of flipping a large number of coins, and average out the result. If we have chosen the generative model properly, it will be guaranteed to converge to the correct answer in a reasonable number of samples. Here is some MATLAB code that runs such an algorithm[4]:

```
clear all
n=10000; %number of samples
ms=zeros(1,n); % a vector to keep track of the wins
t=0.5; %The bias of the coin
```

4 There are more efficient ways of running this algorithm. For instance, we can use a running mean, rather than a large vector. Because the purpose of this section is explanation, simplicity is more important than efficiency.

```
for i=1:n
    k=0; % who is winning: 1 means first player, 0 means second player
    while 1
        k=abs(k-1); % Change who is winning
        if rand<t %Toss the coin
            break %Break out of the loop if it comes up heads
        end
    end
    ms(i)=k; %record the winner
end

mean(ms) %calculate the average
```

A convention is to use about 10000 samples in cases like these. Indeed, when we take 10000 samples in MATLAB, the proportion of times that the first player wins is 0.6639 on one run, and 0.6696 on another run. The true proportion is about 0.6666. We are accurate to about two decimal places.

Why would we use a generative approach, when it is only approximate? Let's consider what happens when we make the game a little more complex. We have a bag of coins, most of which are biased. Each coin has a probability of heads somewhere between 0 and 1, with each value equally likely. We choose a coin at random, and play the game. Does the first player still have an advantage? Is it more or less of an advantage? This is less intuitive. If we try an analytic approach, we end up with the following equation:

$$P(First\ Player\ Wins) = \int_{\theta=0}^{1} \left(\sum_{n=0}^{\infty} \theta(1-\theta)^{2n} \right) d\theta$$

We have an infinite series inside of an integral. We can see already that solving this analytically is not going to be a pleasant experience. Fortunately, we do not need to

do this at all. We can just change a line of code in our generative sampler: Every time we play the game, we randomly sample the probability of heads from a uniform distribution. The one line that was changed, has been highlighted.

```
clear all

n=10000; %number of samples
ms=zeros(1,n); % a vector to keep track of the wins


for i=1:n
    t=rand; %Sample the bias of the coin
    k=0; % who is winning: 1 means first player, 0 means second player
    while 1
        k=abs(k-1); % Change who is winning
        if rand<t %Toss the coin
            break %Break out of the loop if it comes up heads
        end
    end
    ms(i)=k; %record the winner
end

mean(ms) %calculate the average
```

In a few seconds, the sampler generates our answer: about 0.69. The qualitative answer is that the first player has slightly more of an advantage. We can make things prohibitively difficult for an analytic solution, by adding conditions. For instance, what is the probability that the first player wins, given that the game goes for at least three tosses? Coupled with the wrinkle that we select a coin at random from the bag, this makes an analytic solution especially difficult. For example, the probability of going for three tosses is not independent of theta, so we cannot deal separately with these parts of the equation. But in our generative model, all we need to do is add another few lines of code:

```
clear all
```

```
n=10000; %number of samples
ms=zeros(1,n); % a vector to keep track of the wins


for i=1:n
    while 1 %We're going to wait until we have at least three tosses
        t=rand; %Sample the bias of the coin
        k=0; % who is winning: 1 means first player, 0 means second
player
        counter=0;
        while 1
            k=abs(k-1); % Change who is winning
            if rand<t %Toss the coin
                break %Break out of the loop if it comes up heads
            end
            counter=counter+1;
        end

        if counter>2 % if the game went at least three tosses
            ms(i)=k; %record the winner
            break % and go to the next sample
        end
    end % if the game did not go three tosses, try again
end

mean(ms) %calculate the average
```

This takes us a few seconds to run. It spits out the answer of about 0.44. By contrast, a person who tried to solve these problems analytically, would probably still be working on the second problem.

Overall, generative models are easy to work with, and much more flexible than analytic approaches. Rather than solving difficult new equations at each turn, we can just modify the sampler. Generative sampling is especially appropriate when a model needs to be flexibly and repeatedly applied, and exact numerical precision is not important beyond a certain achievable point. This is true of the human data we see in modeling causal reasoning – human data is already so noisy that samplers accurate to two or three decimal places are usually more than sufficient. One strategy, which we

have used here, is to use analytic solutions to a simple version of a problem, to check the validity of a generative approach. Then, we can extend the generative approach with some assurance that it is working properly. This chapter will often use such an approach.

Generative models and language

Language is one place within cognitive science where generative models have been useful. Chomsky (1957) for instance, proposed a set of rules out of which syntactically correct sentences could be constructed: A sentence was correct if and only if it could be built from those rules. More recently, Probabilistic Context Free Grammars (PCFGs) (e.g. Chater & Manning, 2006) have been used to assign a probability to each possibility. Some rules are permitted but rare, which explains why sentences like 'The old man the boats' are grammatical but awkward: their only grammatical interpretation has low probability.

This section brings up language only to show that generative models have a precedent for being successful and tractable models in cognitive science. This thesis does *not* defend any analogies between language and causation. For instance, there is no particular reason to think that representations of causation have morphology, syntax, or phonetics. Neither will this thesis contend that the way that children develop linguistic competence necessarily has anything to do with how they develop representations of causal structure.

Nevertheless, edge replacement is a 'grammar,' but only in the formal mathematical sense: a set of rules for constructing elements of a hypothesis space. To avoid any possible confusion, subsequent sections will refer to the model simply as 'edge replacement' rather than an 'edge replacement grammar.' There *is* however a deliberate analogy to recent work by Charles Kemp and collaborators (e.g. Kemp & Tenenbaum, 2008) on using node replacement grammars to model the development of categorization and other structured learning. Edge replacement is inspired by this work.

An intuitive overview of edge replacement

Like a generative model of language, or coin flipping, edge replacement describes a set of rules for making elements of the hypothesis space. In this case, the hypothesis space is a set of Causal Graphical Models. The way the rules are used defines the probability of each element. This section will provide an intuitive overview of the generative process. Subsequent sections will provide formal details, including an explicit algorithm for generating graphs. A visual overview is shown in Figure 13.

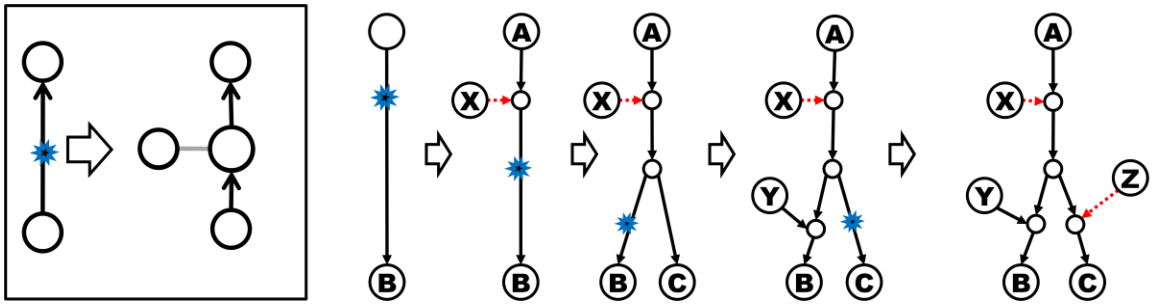

Figure 13: A visual illustration of edge replacement. The box shows the general rule for edge replacement: An edge is replaced with an edge-node-edge path that incorporates the influence of a new node at a specific point.

Edge replacement begins with a single edge between a cause and an effect of interest. The edge has a *length* of 1. The meaning of 'length' here is not a direct analog to physical length – pending discussion later on, we will treat length as a purely formal, functional construct. Edge replacement moves down the initial edge, randomly generating replacements. Each edge replacement incorporates the influence of a new node. To perform a replacement, we first replace the edge with an edge-node-edge path that has the same length. We call this middle node the 'bridge node.' Then we add a new edge (of randomly determined length) connecting the bridge node and a new node.

Each replacement is randomly determined to be one of three types: If it is an inhibitory replacement, then causation at this point follows an AND NOT relation: the effect will be on only if the cause is on, and the new node is off. For instance, in Figure 13, X is generated by an inhibitory replacement. This might be a failure in the electrical system of the car. (These specific cover stories are not generated by edge replacement; they are only used to illustrate the principles involved.) Replacements can also be outward and generative: For instance, C is generated next – C is active whenever the cause is active. We call these 'side effects': for instance, your lights might dim when you start your car. The third type of replacement is also generative, but causation flows inward rather than outward. We call these 'alternative generative causes,' because they follow an OR relation: the effect is active if the cause, or the new node is active, or both. For instance, see Y in Figure 13: this might represent the fact that you can start your car

in other ways, like using a remote car starter. In principle, outward inhibitory replacements are also possible, but are usually irrelevant – we only need to include them for the sake of formal completeness results discussed below. When the new node (not the bridge node) is generated, it is assigned some random probability of firing – this is the only source of randomness in the causal structures defined by edge replacement. Edge replacement is committed to determinism, in the weak sense that variability arises from hidden causes, not from intrinsic randomness in causal relations.

After a replacement, the same process continues along the old path, and along the new edge created by the replacement. Graphs can become arbitrarily complex if replacements are common enough. The process eventually stops when it reaches the end of all edges, yielding a graph. We can 'run' the graph by deciding (again, randomly) whether each exogenous node is on, then propagating causation deterministically through the graph. Note that once the graph has been created, length does not play a role in the determination of which effects are active given the causes. Length is a new concept, but plays a role only in the generative process. Graphs generated by edge replacement are fully compatible with other CGMs, (even those generated in other ways) and subject to exactly the same rules, once they have been generated.

This generative process tends to create a characteristic type of structure called a 'causal stream.' Because of this branching structure, it is helpful to think of causation as flowing down a stream from causes to effects. The remainder of the thesis will use the

terms ‘upstream,’ and ‘downstream,’ to refer to points which would be ancestors and descendants, respectively, if a replacement occurred at the that point in the graph. We will also use the term ‘causal power,’ as a metaphor for the entity that ‘flows’ through the graph. The present use of the term ‘causal power’ builds on the term introduced by Cheng (1997). Thinking of these graphs as pipes, through which causal power flows, and can be blocked and introduced, will yield helpful intuitions. Most of edge replacement’s implications can be intuitively understood by trying to build causal structures out of Y-shaped, but never V-shaped, pipes.

Formal description of edge replacement

This section provides the formal details of how edge replacement operates, in the form of an explicit computational algorithm. A graph with n nodes is made up of three things:

1. An $n \times n$ matrix G that represents the edges. If $G_{ij} = 1$, then a generative edge exists from node i to node j. A zero indicates no edge, and -1 indicates an inhibitory relation.
2. An $n \times n$ matrix L of edge lengths.
3. A vector S of length n that encodes the spontaneous activation probabilities of each node.

Together, the set $< G, L, S >$ comprises a graph. The generative process begins with an edge of length 1 between two nodes (A causes B). We perform replacements by moving along each edge and generating replacements according to a Poisson process

with rate λ, and average length γ, with the restriction that $\lambda\gamma < 1$. This is done as follows:

1. Sample x from $Exponential(1/\lambda)$. If $x > L(A,B)$, then stop. Otherwise do a replacement at point x:
2. Create a new bridge node M as the $(n+1)th$ node.
3. With probability ρ, designate a previously generated non-bridge node as E, choosing each such node with equal probability, otherwise create a new node E as the $(n+2)th$ node.
4. Set $G(A,M) = 1$ and $G(M,B) = G(A,B)$.
5. Set $L(A,M) = x$ and $L(M,B) = L(A,B) - x$.
6. If E already exists, and it is exogenous, set $G(E,M)$, and if it is endogenous, set $G(M,E)$. (This rule is in place to prevent cycles – see below.) Otherwise, with equal probability, choose to set either $G(E,M)$ or $G(M,E)$. Set this relation as −1 or 1 with equal probability.
7. Generate $L(M,E)$ or $L(E,M)$ from $Exponential(\gamma)$.
8. Set $S(n+1)$ = 0 (bridge nodes never activate spontaneously) and generate $S(n+2)$ from Beta(α,β).
9. Set $G(A,B) = 0$, eliminating the original edge.
10. Initiate two new processes along MB and ME, by returning to step 1.

Once edge replacement has stopped, we can use the following procedure to generate an instance of causation on the graph: sample the value of each exogenous node from a discrete uniform using S. That is, node i has probability S_i of being 'on.' Subsequent sections will sometimes refer to a node that is in this state as a node that has 'fired.' Then propagate this activation through the graph to determine the status of non-exogenous nodes, as follows: Each node is 'on' if and only if it has at least one active generative parent, and zero active inhibitory parents.

Properties: Completeness, Validity, and Stopping

Having laid out the rules for edge replacement, the next step is to ensure that the rules do what they are supposed to do: make a coherent set of rules through which we can generate graphs that represent causal relations. This section will argue that edge replacement has three specific properties. First, that edge replacement can be used to represent any given causal-functional relation (completeness). Second, that given these set of rules, edge replacement will never generate anything other than a valid CGM: It makes only directed, acyclic graphs (validity). Third, that the process is guaranteed not to go on forever (stopping).

Completeness

In order to interpret the meaning of edge replacement, it is necessary to distinguish between what Yuille and Lu (2007) call a *causal-functional relation*, (which will sometimes just be called a 'relation' here) and a graph. Causal-functional relations are

properties that exist in the world – facts about the statistical dependencies, direct or indirect, between variables. For instance, in the car example, we described a relation between a set of variables (turning the key, car starting) and then described a series of graphs that each captured the relation in a different way. There exists a one-to-many mapping between relations and graphs. Even under the restricted rules of edge replacement, there may exist multiple ways of constructing any given relation. We can be properly Bayesian and say that the probability of a relation under edge replacement is the sum of the probability of all of the graphs that represent it. Relations that are difficult to construct, in the sense that few or low-probability graphs represent them, have a low prior probability under edge replacement.

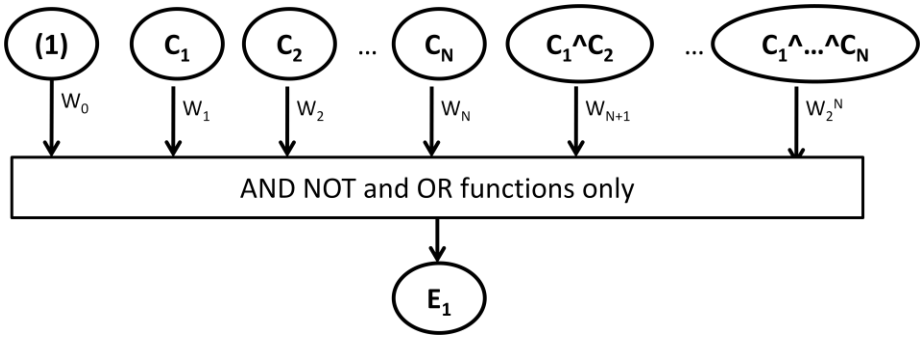

Figure 14: An overview of Yuille and Lu's noisy-logical graphs.

'Completeness' in the sense meant here refers to the idea that there exist no relations that have zero probability under edge replacement. That is, there is always at least one way of constructing any given relation. Happily, Yuille and Lu (2007) have already done a great deal of work on this topic. They define a class of logical graphs called *noisy-logical graphs*, of which the graphs made by edge replacement are a limit case. The schema for a noisy-logical graph is shown in Figure 14. Broadly speaking, the

graph has three layers: first, define the conjunction of every possible set of causes, including the set of no causes (a cause that is always on), and the conjunction of all causes. This creates $2^N + 1$ *factors* for N causes. Assign a noise term to each factor. Then, take the output of these factors, and combine them using any logical function involving just OR and AND NOT. Yuille and Lu prove that this can be used to create any causal-functional relation.

Edge replacement can piggy-back on this result, if we show that edge replacement can mirror any noisy-logical graph. Intuitively, this should not be difficult, because edge replacement graphs are essentially noisy-logical graphs without the noise. We can introduce relations that are formally equivalent to noisy relations, by using hidden inhibitors. Readers who accept this intuition are encouraged to skip the next paragraph, which is less intuitive but more formally rigorous.

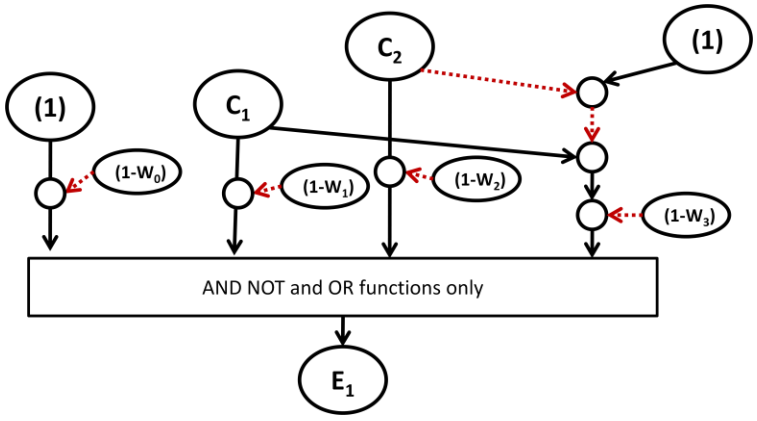

Figure 15: How we can construct graphs equivalent to noisy logical graphs, using edge replacement.

This paragraph will work through each of the three layers, showing how each can be constructed using edge replacement. First, we must create all the factors: every possible conjunction of causes. To do this, first construct a hidden generative cause of

the end effect, which has probability 1 of being active. Then construct each factor of one cause using an inward generative replacement along this edge. To create a conjunction of k causes, start with the first cause, making a generative replacement that connects to it. Then make an inward inhibitory replacement along that generative edge, to a hidden cause that is always on. Then make an inward inhibitory replacement along this inhibitory edge, to the second cause (if this cause has already been used, then ρ must be greater than 0.) Repeat as necessary to add causes to the conjunction. The logic is that the first cause generates the factor, but is prevented by a set of k-1 preventers that are always on. Causes 2 through k prevent their respective preventers from preventing the factor from being on. An example with k=2 is shown in Figure 15. The next two layers are more straightforward. Mirror each noise term by making an inhibitory replacement as the last replacement on each factor, with probability $1 - w_n$ of being on. Finally, the factors can easily be combined in any logical way, because the rules of edge replacement allow for the same OR and AND NOT relations on which Yuille and Lu rely.

Note that this is not the only way of making any given relation edge replacement. This is just a mathematical demonstration that for any given relation, no matter how strange, there exists some representation that edge replacement can make, that will capture the relation. Many relations can be constructed using much simpler graphs, while some relations remain difficult to construct. As we will see later, edge

replacement makes strong, testable commitments about which relations people should find likely a priori.

Because noisy logical graphs can be used to mirror any causal-functional relation, and edge replacement can be used to mirror any noisy logical graph, edge replacement can be used to mirror any causal-functional relation. This shows the completeness of edge replacement.

Validity

Valid CGMs are directed, acyclic graphs. That is, every edge is directed, and there are no paths from a descendant to any of its ancestors. It is important that edge replacement generate only valid CGMs. It is easy to see that edge replacement will never introduce any undirected edges, because no rule creates them. Acyclicity is harder to see. To see why edge replacement cannot introduce cycles, assume for a moment that we did introduce a cycle through edge replacement. This means that at some point, we would have had to introduce a path from an existing node to one of its ancestors. The only time that edge replacement creates new paths from existing nodes, is when we are allowed to re-use nodes ($\rho > 0$). Rule 6 specifies that if an existing node is already endogenous (i.e. if it has ancestors) then any replacements that use that node must point outward. The rules ensure that new paths are either from nodes that have no ancestors, or to nodes that have no descendants. This makes it impossible to introduce a new path from an ancestor to a descendant. Edge replacement cannot introduce cycles.

Stopping

At first glance, there is reason to suspect that the edge replacement process might go on forever. This section will show that the rules of edge replacement strictly prevent this, because λ (the replacement rate) multiplied by γ (the expected length of new edges), is strictly less than one. We can reduce edge replacement to a simplified process in which we increase and decrease the amount of unreplaced edge length, which we will call r. Consider a process in which we start with an edge of length 1, then move down the edge, making replacements. Reaching the end of this edge, we repeat, but only if new length was introduced. The expected number of replacements is λ, and the expected new length per replacement is γ. Therefore, we can capture the process of creating new edges, with respect to r, as:

$$E[r_n] = E[r_{n-1}]\gamma\lambda, \quad where\ r_0 = 1.$$

Repeating this process indefinitely, the expected remaining edge length is:

$$\lim_{n\to\infty} (\gamma\lambda)^n = 0, \qquad since \quad \gamma\lambda < 1.$$

Because the expected remaining unreplaced length approaches zero, edge replacement is guaranteed to stop.

Some extensions and simplifications

Now that we have seen that edge replacement is a complete, coherent, and valid way of generating representations of causal relations, this section will discuss extensions. These will allow us to fit a wider range of data, and increase the range of phenomena on which edge replacement can make predictions. The first extension is branching, a second replacement process through which we can generate multiple tokens of the same type of cause. The second extension will allow us to do inference about causal variables that change over time. The final extension is actually a simplification, through which we can do inferences about classes of graphs that are identical with respect to relevant structure.

Branching

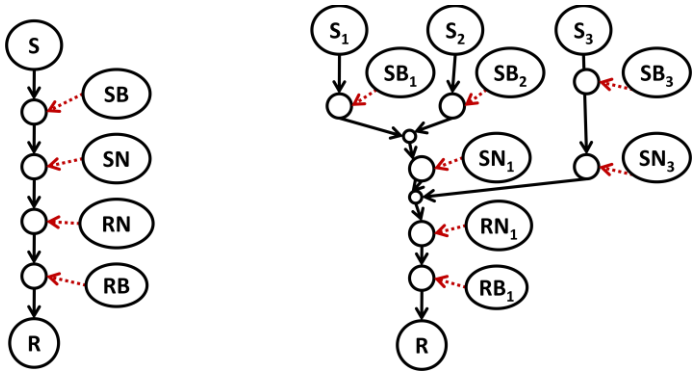

Figure 16: Branching in the cellphone example

The model so far describes causation at the type level; branching will allow us to describe causation at the token level as well. To illustrate this distinction, imagine a network of cellular phones: a call from a friend will cause your phone to ring. Because you have multiple friends, there are multiple instances of this relation. How should we

construct these relations? Intuitively, you do not have a separate phone on which each friend can call you. Rather, there is some point at which all the relations converge on your phone. An intuitive example is shown in Figure 16. At the type level, the causal relation between a sending phone and a receiving phone is inhibited by a the following possible problems: The receiving network is down (RN), the receiving phone's battery is not working(RB), the sender's network is down (SN), or the sender's battery is not working (SB). This expands to the more complex graph shown on the right of Figure 16, when we consider multiple tokens. This graph has some useful statistical implications. For instance, if friend S_1 can't reach you, then you are more likely to have trouble receiving a call from friend S_2 than friend S_3. This is because friends S_1 and S_2 share more inhibitors than friends S_1 and S_3. Branching describes a formal process through which we can make graphs like these.

Like edge replacement, branching is best described intuitively, because the formal algorithm seems overly complex otherwise. Branching works as follows: Once the edge replacement process has stopped, start a new process, with branching rate b and parameter c, such that $b(c+1) < 2$. Generally, b will be lower than λ that is, there will be a fewer, longer, branching events, than replacement events. As with replacement, the branching process moves down the edges, as a Poisson process with rate b. Each branching event makes several copies (on average, $c+1$ copies) of everything either upstream or downstream from the branch point, then attaches these copies to the graph. The copies include all the replacements made on the path, including inhibitors

and alternative causes. We then repeat the branching process on all the other edges, and on any new branches we have made.

We can see branching at work in Figure 16. We choose randomly to start at the receiving cellphone, and move upstream. We make a branch at a random point between SN and RN. We make one extra copy of everything upstream from this point. This leaves us with two branches, on which we repeat the branching process. One more branch occurs between SB and SN, on one subranch. We are left with two cellphone networks, one of which serves two of our friends, and another of which serves one of our friends. The causal structure of each branch is similar, but activation of specific causes are statistically independent. For instance, while both networks can fail, a failure in the left network will disable phones 1 and 2, while leaving phone 3 unaffected.
There are of course some technical wrinkles to iron out with branching. These are dealt with in the full formal description:

Branching on a given subgraph G, moving on a certain path L between C and E, starting at position zero, and moving in a given direction (forward or backward), works as follows:

1. Generate a number from $Exponential(1/b)$ and add it to the current position. If the resulting position is greater than the length of L, then stop.
2. Make a branch at the current position l, which is at length x between nodes A and B:

a. Choose a number of branches from $1 + Poisson(c)$. Make that many additional copies of G, where G_0 is the original, and $G_1 \ldots G_n$are the copies. (Copies of A and B are called A_n and B_n, while the originals are called A_0 and B_0.)

b. Replace the edge on which the branch occurs with an edge-bridge node-edge combination of the same length: Introduce a new node M as the (n+1)th node. Set $G(A_0, M) = 1$, and $G(M, B_0) = G(A_0, B_0)$, $L(A_0, M) = x$, and $L(M, B_0) = L(A_0, B_0) - x$.

c. Eliminate the AB edge on $G_0 \ldots G_n$: Set $G(A_n, B_n) = 0$ and $L(A_n, B_n) = 0$ for all n.

d. Connect the new nodes to the bridge node: If moving forward, set $G(M, B_n) = G(M, B_0)$, and $L(M, B_n) = L(M, B_0)$ for all $n > 1$. If moving backward, set $G(A_n, M) = G(A_n, M)$, and $L(A_n, M) = L(A_0, M)$ for all $n > 1$

e. Portions of the graph that are not connected to G after this step, can be eliminated.

f. Initiate recursive branching: Apply this algorithm, starting at step 1, to all paths $ME_{1\ldots n}$ (if moving forward) or $C_{1\ldots n}M$ (if moving backward). The ith branching process inherits the subgraph $G_i$that remains after step e, and the direction of the parent algorithm.

3. Return to Step 1.

To run branching on a given graph, initiate this algorithm on the main path of each replacement (that is, each path introduced by a replacement) that does not re-use an existing node.

Why branching is guaranteed to stop

Just as edge replacement is guaranteed to stop, so is branching. Consider a process in which we move down a given path of length x, creating branches with rate b, until we reach the end of the path. The expected number of branching events is bx, the expected number of new branches per event is c+1, and the expected amount of new unbranched length per branch is $\frac{x}{2}$ (because each new branch replicates the remaining length, and events are uniformly likely throughout the graph). These expectations all sum and multiply, such that moving a branching process with length x is expected to create $\frac{xb(c+1)}{2}$ of new unbranched length. Because $b(c+1) < 2$, the expected new length introduced by braching on length x, is strictly less than x. As we repeat the branching process, the expected amount of unreplaced length approaches zero ($\lim_{n\to\infty} x\left[\frac{b(c+1)}{2}\right]^n = 0$). Therefore, branching on any path stops. We branch on each path created by a replacement, and the previous argument that edge replacement stops, guarantees that there are a finite number of replacements. Therefore, branching stops.

Introducing time

In the real world, causation happens not just in space but in time: you often act on or observe the same system multiple times, where each instance happens after the previous one. Many of edge replacement's distinct predictions are about what people should expect in these situations. This section will describe the formal infrastructure for incorporating time into edge replacement. We will use the idea of a hidden Markov model.

Hidden Markov models are widely used in probability and statistics. As an example, we can return to coin flipping. Your opponent has two coins: one which is biased towards heads, and another which is biased the same amount towards tails. You and your opponent play a game where you bet on the outcome of the toss. To keep things interesting, at each instance of the game, there is a ten per cent chance of changing coins. The change is hidden, and the coins each look the same, so you don't know when or if the coins have been switched. In the long run, heads and tails are equally likely in this game. However, the trials are not independent: for instance, heads is more likely given that the previous toss was heads. This is because the heads gives us information that raises the probability that we have the coin that is biased towards heads. There is still a small probability that we are working with the coin that is biased towards tails – the coin might have been switched, or the coin might have produced a heads by chance last trial. Uncertainty about the state of the coin induces a dependence between the trails. If, however, we know the value of the hidden state, the trials

become independent. This is due to the Markov condition (which is part of the reason for the name): we know the value of the parents. In a hidden Markov model, each hidden state depends only on the hidden state before it, and the observed states depend only on the hidden states.

We will introduce time to edge replacement by having the values of the hidden causes depend on the values of the hidden causes at the previous time step. Recall that each new exogenous node has an activation probability a that is generated from $Beta(\alpha, \beta)$. To derive the transition probabilities we are interested in, consider a continuous process that works as follows: With probability a, select the staring state to be 'on.' Otherwise, the starting state is 'off.' Then initiate a Poisson process with rate $v > 0$, through time from the starting time (0) to the time of interest (t). At each event, re-sample the state of the system from a.

The parameter v is designed to represent the *volatility* of the system. The lower v is, the fewer change events will occur, and consequently the more the state of the system at each time step will depend on the previous time step. At the limit when v is infinite, each time step will be completely independent – whether the state is on or off will depend only on a.

Another way of thinking about this is that the system switches between two states, using two Poisson processes. One captures the rate at which the system switches off when it is on, and another captures the rate at which is switches on when it is off. The proportion of these two rates determines a, and multiplying both rates by the same

number is the same as increasing the volatility. The two processes (resampling from a at rate v, or two poisson processes with different rates in proportion a) are equivalent, in the sense that some parameterization can be found with the same properties in each case. However, the process described above (using only a and v) is preferred, because the constant rate is easier to work with.

In practice, we will mostly want to work with a discrete-time Markov chain, like the coin example. The underlying continuous model is constructed to make it straightforward to derive a transition matrix from a and v, for any discrete time step t. Consider the following transition matrix:

$$\begin{bmatrix} (1-p) & p \\ q & (1-q) \end{bmatrix}$$

The probability of moving from off to on is p, while the probability of moving from on to off is given by q. Together, t, a and v uniquely determine p and q:

$$p = a(1 - e^{-vt})$$

And

$$q = (1 - a)(1 - e^{-vt})$$

To check that the state has the desired properties, note that the transition matrix above has the steady state $[a \quad (1-a)]$ when:

$$a = \frac{p}{p+q}$$

And

$$\frac{p}{p+q} = \frac{a(1-e^{-vt})}{a(1-e^{-vt}) + (1-a)(1-e^{-vt})} = \frac{a}{a+1-a} = a$$

There are some convenient implications of this formulation. First, because the underlying process is continuous, we can sensibly adapt the model to modified scenarios that discretize time differently. For instance, a discrete-time process where each time step is 10 seconds, can be adapted to a process in which each time step is 20 seconds, by doubling t, which doubles the exponent.

Another implication is that this formulation avoids counterintuitive properties, such as extreme switching behavior, that sometimes become available when transition probabilities are unrestricted. For instance, consider the coin game from the first part of this section: coins with different biases are sometimes randomly selected, but the total probability of heads in the long run is still 0.5 because the biases cancel out. Imagine if the following pattern appeared: the probability of a heads ten tosses from now, given a tails on this toss, is 0.9. While such dependence is formally coherent under an unrestricted Markov chain if we set $t = 10$, and $p = q = 0.9$, it makes no sense given the underlying process. To see this, consider the fifth toss. Is it more likely, or less likely, to be heads, than 0.9? The process would have to know what toss it was on. To phrase this more formally, the underlying Poisson process is *memoryless:* given a known rate, the probability of an event occurring at a given time is independent of the number of events that occurred before. This makes it impossible to construct discrete transition matrices in which previous transitions make subsequent transitions less likely. Such transitions are necessary in order to create extreme switching behavior.

This does not mean that edge replacement is committed to the idea that causal systems are always memoryless in time. Far from it. It only means that any time-memory should be modeled in the graph, not offloaded onto the function that generates the activation of the exogenous node. For instance, the transition probability of political parties in power could reasonably be more extreme than their proportion of long-run victories. This could be due to voter's tendency to be especially critical of the party in power. In this case, the graph that gave rise to that relation could not model the node 'Party A in power' as an exogenous node under edge replacement. Rather, it would have to specify a process through which being in power caused election loss. It would have to represent, at least tentatively, the reasons why this process occurs as it does, for instance in the form of hidden inhibitory nodes.

Another way of explaining this point is to say that there are two ways for edge replacement to capture time: One is through the parameterization of individual hidden causes (which must be memoryless), another is through specifying events that occur at different times using different nodes (which need not be memoryless). This is consistent with edge replacement's general approach: Variability that cannot be expressed using a restricted set of simple probability distributions, should be explained via causal structure, not a complicated probability distribution.

Simplified Edge Replacement

Another extension to edge replacement will make it not more complex, but simpler. In some cases, we are not interested in the specific intermediate causal structure that

gives rise to a particular relation; all we are interested in is the total probability of all the possible intermediate structures that give rise to a relation. This is much like a case in which we are concerned with the probability of having five heads in a series of coin tosses, without regard for the specific order in which the tosses occurred. This section will describe a computational shortcut for doing inference over all possible edge replacement-generated graphs that capture a given relation. This particular shortcut will allow us to model nonindependence effects using fewer parameters than would be necessary if we used the full model. It will be referred to subsequently as *simplified edge replacement.*

Consider a causal relation between C and E. For most relations, there exist multiple structures that edge replacement can generate, that give rise to the relation. If we are not interested in the specific causal structure that gives rise to the relation, or in the status of other events than C and E, then only two things matter, from edge replacement's point of view: First, it matters whether an active, inward replacement occurs in which the causal power from the replacement actually reaches the main path connecting C and E. We will call this a *relevant* replacement from now on. If no such replacement exists, then no matter what the intermediate structure is, the end effect will be the same at the main cause. Second, it matters what the status of the *latest* relevant replacement is. Figure 17 shows an example.

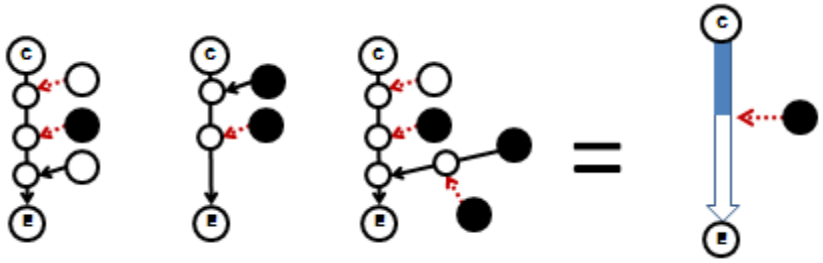

Figure 17: Examples of simplified edge replacement.

All three graphs have the same latest relevant replacement, which is an inhibitor. Note that the graph on the right, the causal power from the inward generative replacement does not reach the main path, because it is inhibited along the way – thus it is not relevant. The three graphs are equivalent with respect to simplified edge replacement. We will represent this whole class of graphs using the visual shorthand on the right, in which open arrows represent the class of paths which are equivalent with respect to their latest relevant replacement. Blacked in nodes represent nodes that fire on a particular instance, while open nodes indicate nodes that did not fire.

The parameter h captures the rate of relevant replacements. We move down the edge making relevant replacements according to a Poisson process with rate $-\ln(h)$. This means that the probability of having zero relevant replacements on a path of length 1 is equal to h, and h ranges from 0 to 1. In general, on edges of length l, the probability of at least one relevant repalcement is $1 - h^l$. Because we assume that generative and inhibitory replacements are equally likely, the presence of at least one relevant replacement means that the probability of the effect occurring is 0.5. Therefore, the probability of the effect being generated by this edge is:

$$p = ch^l + 0.5(1 - h^l)$$

Where c is the status of the main cause (1 if it is on, zero otherwise.) We can combine this with a, the probability that the effect activates spontaneously, using a Noisy-OR:

$$P(E = ON) = p + a - pa$$

This equation is a simple way of doing inference over all possible intermediate structures, when the specific intermediate structure is unknown and does not matter. Note that simplified edge replacement does not make different predictions than edge replacement – it is exactly equivalent to edge replacement, but marginalizes out certain inferences about intermediate structure. This is similar to the coin example again: we will get the same answer by considering all possible ways of getting five heads out of ten, as we will considering each case one by one and adding them up. When some intermediate structure *is* known, we can still use simplified edge replacement on subparts of the model for which we do not know about intermediate structure. We will see examples of this below, when fitting data about nonindependence.

Overall, simplified edge replacement allows us to reduce the parameters of the model: λ, γ, α, and β,[5] collapse into h and a. That is, some setting (possibly multiple settings) of λ, γ, α, and β corresponds to each setting of h and a. Despite the fact that it does not explicitly represent intermediate structure, simplified edge replacement makes different predictions from minimality in many important situations, as we will see below. In particular, because simplified edge replacement implicitly assumes that

[5] Simplified edge replacement assumes that ρ , the probability of re-using existing nodes in a replacement, is zero or negligible.

collateral causes and effects are generated from a length-based branching process, in which inhibitors occur at a specific point, it will allow us to fit nonindependence effects.

A note about length

Before proceeding to use edge replacement to fit experimental data, it is important to review the role of length in the model. Length plays a role in many of the model's key predictions, that may not be obvious at first.

Note that length plays a role only in the generative process for a given graph, but not in the process of actually generating instances of success and failure for a given graph. As an example, consider a simple graph that we will call Graph A, with one inhibitory replacement along the main edge, at the location of 0.3, and another graph that we will call graph B, with one inhibitory replacement at the location of 0.5. If we know for certain that this is exactly the underlying structure, then these two graphs are identical once the generative process is complete. The probability of the effect firing, given that the cause fired, will be determined by whether the hidden inhibitor is active.

On the other hand, consider a case in which we know that there are no replacements upstream from the inhibitor, but the presence or absence of further replacements downstream from the inhibitor is uncertain. In this case, Graph A has more length on which further replacements might fall, than Graph B. Other things being equal, we would expect a higher probability that the status of the effect would be determined by the cause, in Graph B than in Graph A. In this case, length plays a role in

the process of inference under uncertainty, because we are uncertain *where* (in the sense of length) replacements might fall.

When we integrate over multiple hypotheses like this, length takes on a probabilistic flavor – more apparent randomness appears on longer paths, than on short ones. True to the Laplacian stance of edge replacement, this apparent randomness is due to uncertainty, not due to the intrinsic randomness of causal relations. Nonetheless, the probabilistic flavor of length plays a role in our predictions under conditions of uncertainty; conditions that hold most of the time in the real world.

Applying edge replacement

We have now developed edge replacement to the point where it can make predictions about the experimental data discussed in chapter 2, and begin to make further predictions. Recall that minimality had difficulty with nonindependence, determinism, and mechanism. We will begin with nonindependence.

Nonindependence

Recall that Walsh and Sloman found a basic nonindependence effect: Participants were told about a causal relation, and asked to judge the probability of the effect given the cause. They were then told that a collateral effect had failed to occur. Participants then lowered their initial judgments of the probability of the effect, in violation of minimality. Probability judgments for the two cases $P(E_1|C), P(E_1|C, \bar{E}_2)$ averaged over Walsh and Sloman's experiments, are shown in Figure 18.

Figure 18: The model class, data and model predictions for Walsh and Sloman (2008)

Model predictions were generated using sampling over the whole set of models defined by edge replacement, that fit the cover story given to participants in Walsh and Sloman's experiments. The program generates a set of graphs as follows: Start with one cause and one effect, with an edge of length 1 connecting them. There were exactly two visible effects, so we need to generate at least one side effect. This is equally likely to have occurred at any given point along the main path, so choose a bridge point M at location $x \sim U[0,1]$, and make the side effect E_2 via a replacement at this point. We can assume that the two effects are equally likely, which implies they should have the same distance on their path from the main cause[6]. For this reason, the sampler set $L(M, E_2) = L(M, E_1) = (1 - x)$. This captures all the visible causal structure – further intermediate structure can be handled using simplified edge replacement. Figure 18

[6] The two experiments below explicitly told subjects that the effects were equally likely; Walsh and Sloman did not. Nonetheless, to maintain consistency, the same program was used to simulate all three sets of experiments. To be sure that this did not bias the model predictions, a simulation was run in which the length of (M, E_2) was generated randomly from $Exponential(\gamma)$, as described in the full edge replacement model. The sampler set γ at 0.5. This produced the same results to two decimal places. Because the extra parameter was unnecessary, this section reports the procedure and results from the more restricted simulation, whose methods were more in line with other simulations.

shows the graph from simplified edge replacement used in modeling Walsh and Sloman. Note that the graphs were not hand-selected for this experiment: the left part of Figure 18 represents *all* the possible graphs that edge replacement defines, that are consistent with Walsh and Sloman's cover story.

For each sample, the program randomly generated a graph, then sampled the value at the end of each edge using simplified edge replacement, with parameters a=0.3, and h=0.5. For example, to determine the status of an end node, first sample the status of the bridge node, then use the status of the bridge node to sample the status of each end node. The simplest case, determining $P(E_1|C)$, can be solved analytically, because the Poisson rates sum to one:

$$P(E_1|C) = p + a - pa, \text{ where } p = ch^l + 0.5(1 - h^l) = 0.5 * 0.5 + 1 * (0.5) = 0.75$$

$$\text{So: } P(E_1|C) = 0.75 + 0.3 - 0.75 * 0.3 = 0.825$$

This is the same value achieved through sampling. The results are shown in Figure 18 along with experimental results from Walsh and Sloman. For instance, the value of $P(E_1|C, {\sim}E_2)$ is the number of samples one which $E_1$was active, and E_2 was not active, divided by the total number of samples in which E_2 was not active. The sampler ran until every category had at least 10000 samples.

As is clearly visible in the graph, the fit is good. Edge replacement naturally predicts a nonindependence effect because inhibitors along the edge CM inhibit both effects, creating a correlation. Furthermore, as Figure 18 shows, varying the parameters did not significantly change the predictions of the model. (Quantitative measures of fit

will be provided below, once more data points, from multiple experiments, can be considered.) Edge replacement explains, and quantitatively predicts, the nonindependence effect that Walsh and Sloman discovered. Given that we have fit only two data points, however, and there are two parameters, this is still a relatively weak test of the model. Fitting further experimental data will strengthen the case for the model.

Fortunately, further experimental data exists in the literature. For instance, recall that Rehder and Burnett (2005) found a similar nonindependence effect when investigating causal inference about features of a category. Recall that they told subjects that one feature caused three other features. In violation of minimality, subjects changed their judgments of the probability that a feature was present, depending on the number of other features that were present. Their data are shown in Figure 19: recall that the staircase pattern indicates a nonindependence effect. Also recall that in Experiment 2, Rehder and Burnett used blank predicates, but nonetheless found a nonindependence effect.

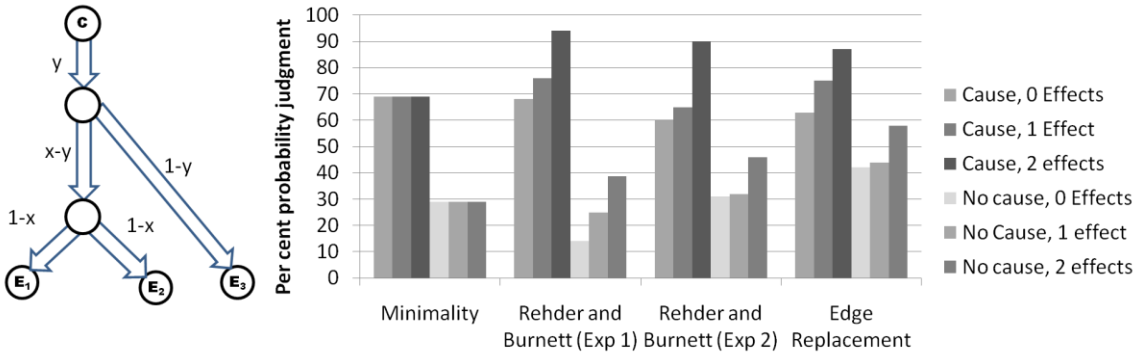

Figure 19: Graph class, data and model predictions for Rehder and Burnett (2005).

Edge replacement's predictions were created by modifying the program used to fit the data from Walsh and Sloman. Instead of two effects, there were three. The third effect was generated by selecting a third bridge point at a second uniformly chosen point. In this experiment, participants were explicitly told that the probability of each effect feature was the same, so the sampler ensured that the path from the cause to each effect was length 1. As is apparent in Figure 19, the graph is not symmetric in this case: one effect (E_3) always had a longer path than the other two effects. To avoid any problems created by the asymmetry, the sampler randomly permuted role of each effect when sampling. Otherwise, the sampling scheme was the same. Figure 19 shows the results, using exactly the same parameter settings as used when fitting Walsh and Sloman. ($a = 0.3, h = 0.5$)

Because Rehder and Burnett also asked subjects to make judgments of the probability of the effect in the absence of the cause, these model predictions are shown as well. Note that in the absence of the cause, the model predicts a slightly higher probability judgment than Rehder and Burnett observed (though the model captures the nonindependence effect they observed). This may be because of the well known *causal status effect* (Ahn, Kim, Lassaline, & Dennis, 2000): Features that cause other features are seen to be more important in category membership. Subjects may have thought that instances lacking this central feature were less likely to be members of the category, and therefore less likely to have the effect features as well. Rehder and Burnett verified this experimentally within their paradigm: Subjects were significantly

more likely to infer category membership when C_1 was present than when it was absent. Other features had an effect on categorization as well, but none as much as C_1. Edge replacement could accommodate this finding by introducing an unobserved category essence as an alternative generative cause of multiple features, with lowest distance to C_1. The absence of C_1 would be diagnostic of the absence of the essence, and lower the probability of the effect features. However, the goal in these simulations was to use as much as possible the same sampler across experiments.

Finally, we can fit the result from Mayrhofer et al. (2010). Experimental results are shown in Figure 20. These numbers are averaged across three similar experiments, two published in Mayrhofer et al (2010) and one from an unpublished manuscript (personal communication from Mayrhofer). Recall that Mayrhofer et al. told subjects about four telepathic aliens: the cause alien could cause the effect aliens to think of food. In the sending condition, they described the causal relation as follows: The cause alien sends his thoughts to the effect aliens, but the cause alien sometimes has trouble concentrating, and fails as a result. In the receiving condition, subjects were instead told that the effect aliens read the thoughts of the cause alien, and each effect alien sometimes fails because of a failure concentrating. Note that in the sending condition, Mayrhofer et al described a single failure early in the causal process, while in the reading condition, they described multiple inhibitors late in the causal process.

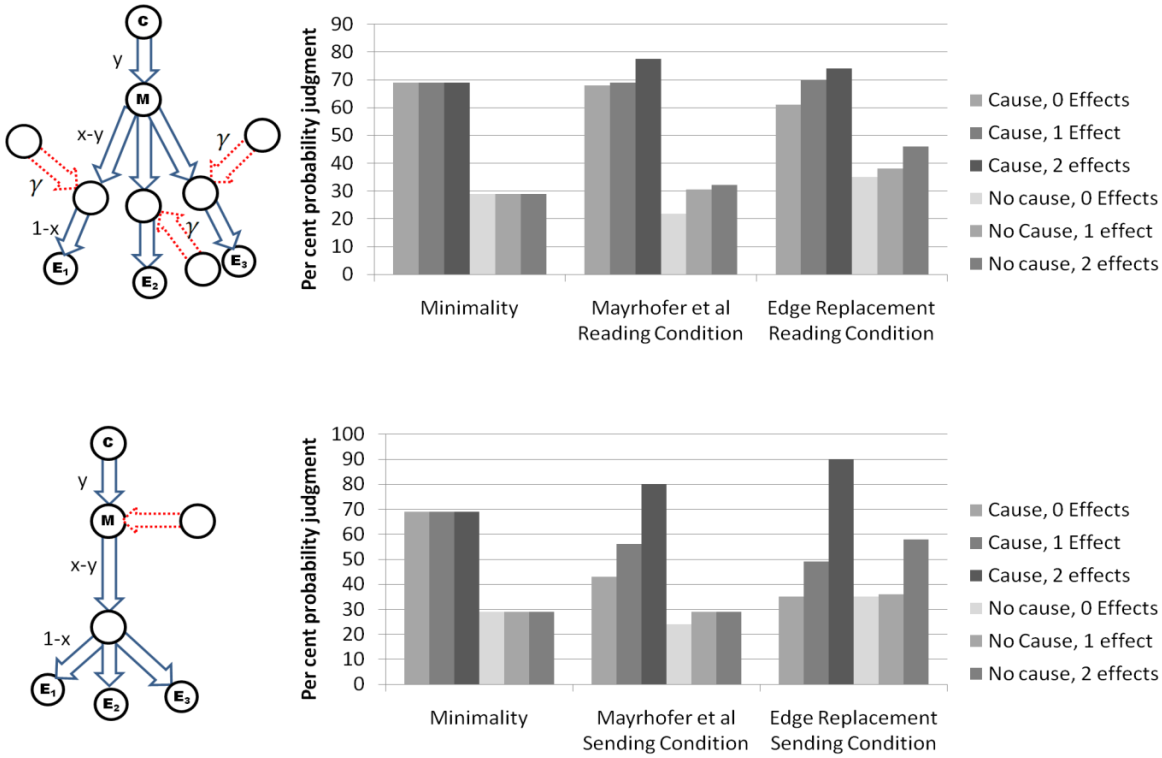

Figure 20: Graph classes, data, and model predictions, for Mayrhofer et al (2010). The top graph shows the reading condition, while the bottom graph shows the sending condition.

The edge replacement sampler was modified to accommodate these cover stories, as follows: First it generated a single cause and a single effect, with one inhibitory replacement at a uniformly chosen point x. The inhibitor was given probability a=0.3 of firing. Because the inhibitor is explicitly described, it cannot be captured using simplified edge replacement, and so the path from the inhibitor to the main path must have a length chosen from $Exponential(\gamma)$. Because setting $\gamma = 0.5$ yielded similar results to simplified edge replacement in modeling Walsh and Sloman's data, the sampler set $\gamma = 0.5$. Because there are three tokens of the same type of effect, the sampler used branching to create three instances of the effect. Note that if and only if the branch

point is upstream from the inhibitor, there will be three copies of the inhibitor. Because the reading condition explicitly states that there are three separate inhibitors, any graphs in which the branch point was upstream from the inhibitor, were assigned to the reading condition. Conversely, graphs in which the branch point was downstream from the inhibitor, were assigned to the sending condition. No other changes were made to the sampler; the rest of the inference used simplified edge replacement.

Note that this is a straightforward application of the rules of edge replacement to the cover story used by Mayrhofer et al. Branching is the correct approach, because there were three similar tokens of the same type of effect. Because a an inhibitor was explicitly described, there was no alternative but in include it in the model. The inhibitor was incorporated using the same assumptions used in modeling other data. The data were not explained using complicated or post-hoc assumptions.

Results are shown in Figure 20. Again, the main parameters were the same as in fitting the other two experiments ($a = 0.3, h = 0.5$). Note that no parameters were varied between conditions. The difference between the conditions was accommodated directly through structural aspects of the model, which were necessary to fit the cover stories. This is in contrast to Mayrhofer et al's model, which introduced a source of common inhibitory noise, linked directly to each effect (see Figure 21). This graph cannot be considered strictly minimal, because it introduces new edges that create correlations between collateral effects, before any data is seen. However, it represents

a simpler graph (in the sense that has the fewer edges) than edge replacement, and thus is closer to minimality than edge replacement.

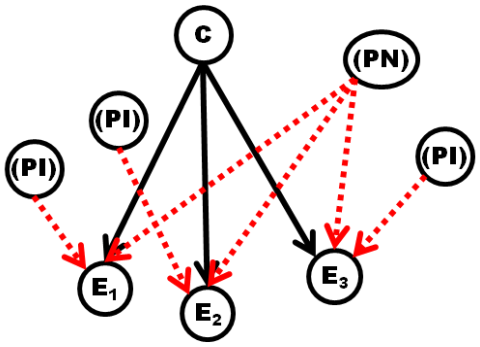

Figure 21: The common inhibitory noise model used by Mayrhofer et al (2010).

Mayrhofer et al's model had two parameters: the strength of individual inhibitory noise (PI), and the strength of common inhibitory noise (PN). They varied the balance between these two parameters between conditions –common noise was set to be strong in the sending condition and weak in the reading condition. Note that Mayrhofer et al's model could have accommodated the converse pattern just as easily (stronger nonindependence effects in the reading condition) and therefore lacks the explanatory power of edge replacement, which is committed to its predictions. A quantitative comparison of the two models is not presented because both provide comparable fits, and use a comparable number of parameters. The difference lies in the fact that edge replacement explains the difference between conditions, rather than just accommodating it.

Edge replacement provides insight into why the description of the mechanism changes the degree of nonindependence observed, in Mayrhofer et al's experiment. The path connecting C and M is shared between all three causes – replacements here create

nonindependence effects. In the sending condition, the replacement described as responsible for most failures (failure concentrating) is described in such a way that it must occur along this path. That is, in the sending condition, the inhibitor is described as *early* in the causal stream. By contrast, the reading condition, the inhibitor is described as *late* in the causal stream. In general, edge replacement predicts that other things being equal, in a branching causal stream with multiple effects, early replacements should create more nonindependence than late replacements. Conversely, edge replacement predicts that in a branching stream with multiple causes of a common effect, early replacements should create less nonindependence than late replacements. Subsequent chapters will refer to this prediction as a *stream location effect.* Chapter 3 will test for stream location experimentally in preschoolers.

Now that we have fit sufficient data, we can evaluate the quantitative measure of fit between edge replacement and experimental data, and compare it to that of minimality. The minimal models shown in Figures 18, 19, and 20 were constructed as follows: Fit one value for the probability of the effect in the presence of the cause, and one value for the probability of the effect in the absence of the cause. This was done by taking the mean judgment in each case. Because edge replacement used the same parameters across experiments, these two parameters for minimal models were also the same across experiments. According to the minimal model, the deviations from its predicted value should be random – they should not be significantly correlated with any model beyond what minimality predicts. However, the values were significantly

correlated with edge replacement's predictions: at the level of $r = 0.95, p < 0.01$ in the presence of the cause, and $r = 0.67, p < 0.05$ in the absence of the cause.[7] The correlations for minimality were undefined, because there is no variance in minimality's predictions. Edge replacement explains a significant amount of the variance that minimality leaves unexplained, without using any more parameters.

Edge replacement explains nonindependence effects in common cause structures. The literature contains further data on nonindependence, which edge replacement can also accommodate. However, an exhaustive review will add little to our discussion. This chapter has focused instead on data that shows basic, foundational nonindependence findings (Rehder and Burnett, and Walsh and Sloman's initial experiments) and on data from which edge replacement makes testable predictions for future research (Mayrhofer et al's stream location effect.) At this point, we turn to a phenomenon that at first glance presents a challenge for edge replacement: causal chains.

Causal Chains

Classically, there are three canonical causal structures investigated in causal reasoning research: The common cause (A causes B and C), the common effect (each of A and B cause C) and the chain (A causes B, which in turn causes C, which in turn causes D, etc.).

[7] Another approach would have been to look at the overall correlation, grouping together the data from the presence and the absence of the cause. This is an incorrect approach, however, because both the experimental data, and model predictions, were grouped into two widely dispersed clusters. Such approaches create artificially high correlations. Nonetheless, when using this analysis edge replacement (r=0.91) still outperforms minimality (r=0.81)

We have already seen that in many cases, when people are told a cover story that suggests a common effect model, they do not represent the minimal common effect model shown in Figure 22a. As we just saw, the data are better fit by a branching stream such as those generated by edge replacement. Conversely, edge replacement predicts similar nonindependence effects in common effect scenarios, which we will investigate experimentally in Chapter 3. Edge replacement also makes commitments about how people should represent causal chains. Edge replacement suggests that just as in the case of a common cause model, people told cover stories that indicate a chain, actually represent something slightly different.

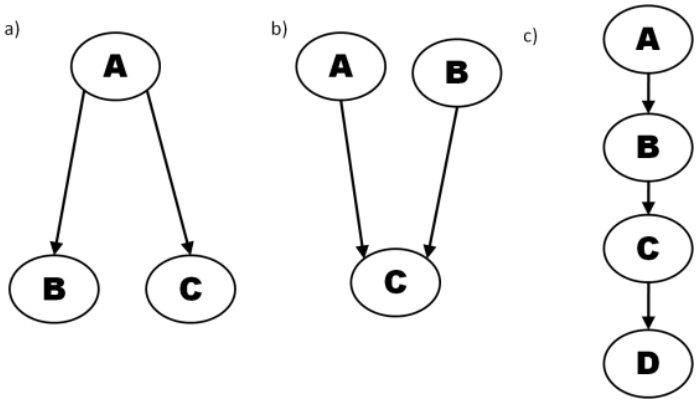

Figure 22: Canonical common cause, common effect, and chain structures.

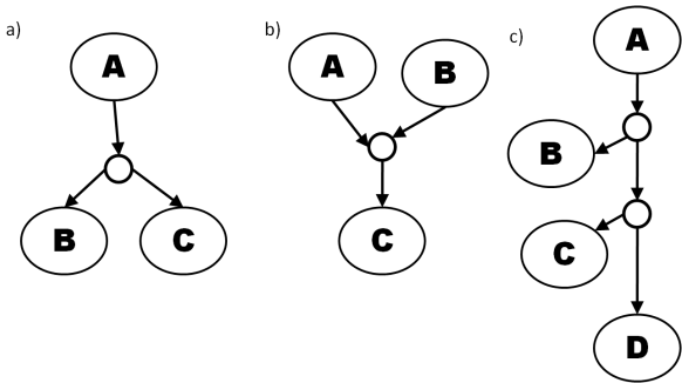

Figure 23: Edge replacement versions of canonical structures.

Just as the canonical common effect model is the limit case of edge replacement's branching, the canonical chain model shown in Figure 10c is a limit case for edge replacement. To see why, recall that bridge nodes are never visible. Therefore, edge replacement cannot represent a chain of three observable events that directly cause each other. Rather, each observable event must be on a branch (even a very short branch) off the main path. If the branches are sufficiently short that they have no replacements on them, then the graph is functionally equivalent to a chain. In the limit case where each branch has zero length, the graph *is* a chain. As an example, imagine you are told the cover story that eating purple jelly beans causes increased heart rate, which causes increased metabolism, which causes weight loss. Contrary to the canonical, minimal model, edge replacement holds that you represent the event 'increased heart rate' as something that could in principle be disabled, without preventing the jelly beans from causing weight loss. For instance, there might be an error in measuring the heart rate, or a countervailing decrease from an alternative cause, that still allows the increased metabolism to occur. Edge replacement predicts

that in a default representation of a causal chain, such possibilities exist, because the branch exists and is vulnerable to replacements. This formal commitment is related to a philosophical commitment that there is always a gap (however small) between a real causal event, and the observation of that event (Hume, 1748/2000).

If we assume that people represent causal chains like Figure 22c, but they in fact represent them as in Figure 23c, then we should see apparent Markov violations when people are reasoning about causal chains. In particular, the minimal, canonical model predicts that given information about events one step away in the chain, information about events more than one step away, should provide no information. For instance, given C, D tells us nothing about A. By contrast, such information is allowed to influence probability judgments if we use the graph in Figure 22c because it is possible that events one step away in the chain were determined by events on the branch, not on the main path. For instance, given that a person's metabolism did not increase, hearing that their heart rate increased, can still rationally raise the probability that they lost weight, because their metabolism might have been modified by irrelevant factors.

Contrary to minimality, and just as edge replacement predicts, multiple researchers have found apparent Markov violations in chain models. These include Rehder and Burnett (2005), and Mayrhofer et al. (2010). For instance, Rehder and Burnett told subjects that feature A caused feature B, which caused C, and then in turn D, in a replication of their previous experiments, this time with a chain structure. As before, they asked subjects to judge the probability of each feature given a set of other

features. Features more than one step away influenced probability judgments, violating the Markov condition. Rehder and Burnett's data are shown in Figure 24 along with predictions from edge replacement, and minimality.

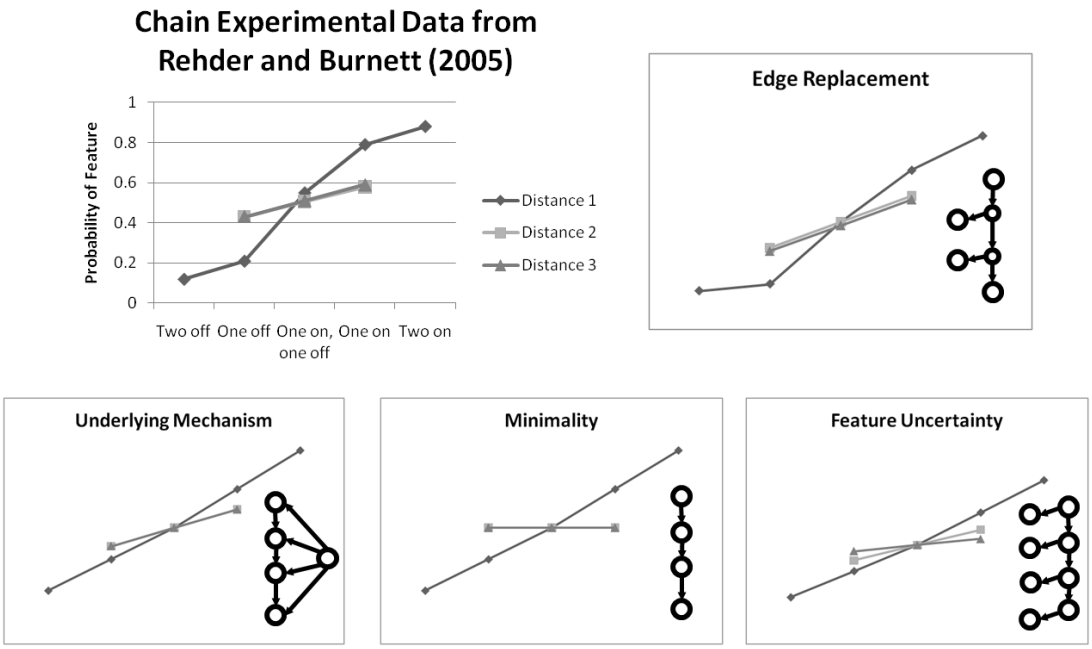

Figure 24: Chain data from Rehder and Burnett (2005), along with qualitative model predictions.

Rehder and Burnett asked each subject about each of 32 possibilities: Whether each of the four features was likely to be present, given each of the $2^3 = 8$ possible configurations of the other features. They them summed these judgments to generate each point in the graph. For instance, the distance-1, two off point, is the average probability judgment summing over all cases in which the judged feature had two neighbor features that were not present, and the distance-3, one on point, is the average judgment over cases in which exactly one feature, three steps away, was present.

Rehder and Burnett compare the experimental data to the qualitative predictions made by three types of models, shown in Figure 24. Note that Rehder and Burnett did not provide quantitative predictions (i.e. their axes had no numerical values), so these cannot be displayed – only the predicted shape is reproduced here. The model labeled 'minimality' is a basic chain as shown in Figure 22c; it predicts that only events one step away should influence probability judgments. (Note that Rehder and Burnett refer to this model as the 'chain model.') The model that Rehder and Burnett call the 'feature uncertainty model' introduces a mediating link between each feature and the observation of the feature. (The structure of the feature uncertainty model is similar to that of edge replacement, but only in the case of causal chains.) Rehder and Burnett's 'underlying mechanism model' is a canonical chain, but with an additional hidden common cause of each event.

Edge replacement's qualitative predictions are included as well, for comparison. These predictions were generated as follows: The sampler started with the edge AD. It then generated B and C using two replacements along this edge, each at uniformly chosen points, with branch length chosen from $Exponential\ (0.25)$[8]. The root cause was on 75 per cent of the time (as specified in the cover story) but the sampler used simplified edge replacement to generate the status of each of B, C, and D, using $h = 0.5$, and $a = 0.3$ as before. To determine a probability judgment for a given question asked

[8] This was lower than the value used in fitting Walsh and Sloman (2008), where $\gamma = 0.5$. The value was lowered to avoid creating a counterintuitive situation in which the path from A to C was expected to be longer than the path from A to D. Other values of γ, including 0.5, yielded similar qualitative predictions.

to participants, the sampler compared the number of times that the feature in question occurred, to the number of times it did not occur, when the other features were held constant.

Edge replacement captures the effects observed in the human data: Events one step away are predicted to have the strongest effect, but effects two and three steps away are predicted to have an effect as well. Minimality does not predict that nodes two and three steps away will have an effect, because such an effect would violate the Markov condition. Edge replacement also captures the S-shaped curve observed in the one-step data. None of the models considered by Rehder and Burnett predict this S-shaped curve. One explanation for this effect is that it is an artifact of the experimental design: Events that have exactly one failure, or one success, one step away, are always either the root cause or the end effect. These nodes are more strongly affected by other events, because they lie on the main path. Overall, edge replacement seems to fit the data as least as well as the other models that Rehder and Burnett consider.

These results are not conclusive; more work is needed to properly develop edge replacement's predictions about causal chains. The purpose of this section was only to show that edge replacement can remain coherent when dealing with causal chains. Just as in the cause of common cause models, the fact that edge replacement is incapable of representing canonical chains, in which each event directly causes the next, may actually be a solution rather than a problem. That is, edge replacement may be closer to the way people represent chain-like causal relations, than the canonical, minimal model.

Determinism

Edge replacement can also explain determinism effects, for instance those observed by Schulz and Sommerville (2006). Recall that children in their experiments observed a series of successes, followed by a series of failures attributable to a visible inhibitor (a ring), and then a period of sporadic failures with no visible inhibitor. Children inferred that a hidden cause (a flashlight) had inhibited the causal relation during the sporadic period, as indicated by their use of the revealed cause to inhibit the effect. No such effect was observed when there were no sporadic failures – children instead chose the ring. Note that this is a qualitative result – children are not making probability strength judgments, but binary choices.

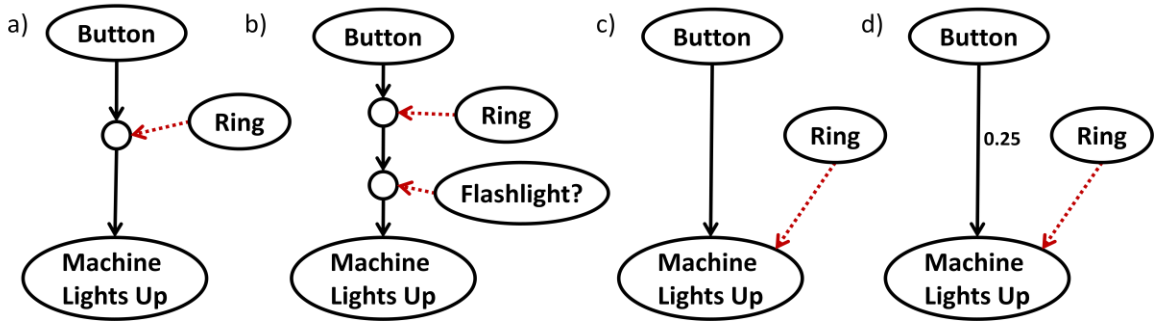

Figure 25: Four possible representations of the causal structure used in Schulz and Sommerville (2006).

We can model this result with edge replacement as follows: Consider a causal relation between a single cause (such as pushing the button) and single effect (such as the machine lighting up) which is disabled by a visible inhibitor (the ring). The simplest edge replacement graph that fits this relation is shown in Figure 25a. For most settings of λ, this type of graph will also be the most likely. Other graphs, for instance those with additional hidden inhibitors, take up the remainder of the hypothesis space. Thus, in the

absence of sporadic activation, Figure 25a shows the graph with maximum probability. If, however, we show children a period of sporadic activation with no visible inhibitor, then Figure 25a is no longer consistent with the data. Figure 25b shows the simplest consistent graph when there is sporadic activation. Again, given the data, this graph is the most likely[9]. It contains a hidden inhibitor, which children apparently map on to the flashlight when it is presented. According to this graph, the flashlight and the ring are equally likely to disable the machine. Children presumably show a slight preference for the flashlight due to its novelty. Recall that children do *not* show this preference in what this section will call the *solid*[10] condition, where there is no sporadic activation (the machine always works, unless the ring is on it), so the novelty preference cannot explain all the experimental results. Edge replacement qualitatively explains not just Schulz and Sommerville's main result, but also the difference between conditions.

Minimality cannot explain this difference between conditions. Minimal structures shown in Figures 25c and 25d fit the solid and sporadic conditions respectively, without positing a hidden inhibitor in either condition. Because minimality allows probabilistic functional forms to accommodate simple graphs, sporadic activation should give no special motivation for children to posit a hidden inhibitor, any more than solid activation.

---

[9] This assumes, of course, that the inhibitor is intervened on by a person, to produce exactly the right pattern of activation for the pattern of data observed in the experiment. Schulz and Sommerville's data suggest that children make this assumption. If instead we were to assume that the inhibitor changed states spontaneously, a more complex graph might have higher probability under some volatility settings.

[10] Schulz and Sommerville call this the "deterministic" condition. This thesis uses the term "solid" to avoid confusion, because edge replacement posits a different kind of determinism.

Edge replacement makes further predictions about determinism, that can be tested in experiments. For instance, edge replacement with time predicts that people should infer the consistency of successes from the consistency of failures. Further detail about these predictions, and experimental tests, will be provided in Chapter 5.

Mechanism

We have seen so far that edge replacement makes sound predictions in the cases of nonindependence and determinism, on data sets that were problematic for minimality. We turn now to a more theoretical question: Causal mechanisms. As discussed in Chapter 2, mechanism is often about intermediate causal structure, about which minimality is conservative: Mechanism is minimized away. Chapter 1 promised that edge replacement would provide a better answer to the question of mechanism – this section will address that question.

Mechanism is a difficult concept to discuss, because it means many things to many researchers. Many of these meanings are vague. Nonetheless, it appears to be an important part of people's causal reasoning. For this reason, mechanism is a good question, but not necessarily a good answer. It is a mistake to rely too heavily on an approach that defines the phenomenon in advance, then seeks to explain it. This approach will only lead to affirming our original conception. Rather, we should be open to the idea that formal models like edge replacement might tell us something new about the scope of the phenomenon of causal mechanisms. The introduction broadly defined a causal mechanism as *events that happen between cause and effect*. While this

definition doesn't do much on its own, it is sufficiently broad to remain open to new insights, without being hopelessly vague.

Within this broad definition of mechanism, there is room for a spectrum of possible commitments about representations of mechanism. Recall the jelly bean example: Imagine that we determine statistically and experimentally that eating purple jelly beans causes people to lose weight. A person accepts and represents this causal relation. How much are the events that happen between cause and effect, an essential part of this representation? Different answers to this question represent different theories of mechanism. At one extreme, we can situate an entirely mechanism-free account: No representation of mechanism is necessary. This is the position endorsed, intentionally or not, by minimality. At the other extreme, we can situate a broad application of force-dynamical accounts of causation: Representing any causal relation necessarily involves imagining physical forces moving through a real concrete causal system; all the physical details of the system must be filled in, if tentatively, in order to represent the system at all. While there are formal accounts of force dynamics that can represent simple physical events such as colliding billiard balls, it is far from obvious how to account for such complex, medical causation using force dynamics. Hand-waving becomes necessary. Intuitively, neither seems right; we are left with a desire to integrate the two extremes. This requires even more hand-waving.

Edge replacement allows us to clearly articulate an intermediate position. The theoretical implications of edge replacement for representations of mechanism, which will be discussed below, are as follows:

1. A cause-effect relation involves the movement of causal power through a functional space, which is grounded in and constrained by physical space and time.
2. Other things being equal, causal power moves through the space in a deterministic way: Variations in the progress of causal power are themselves due to events that have a specific location and extent in the functional space.
3. Some functional arrangements are more likely than others. For instance, causal relations tend to be arranged in branching streams.

These implications begin with the implications of using length in the generative process. Recall that in the formal treatments, length has had a somewhat circular definition so far: Longer edges are just those that tend to have more replacements on them. Now that we have seen the implications of using length, we can interpret its psychological meaning. Length is *functional distance.* This is not necessarily the same as physical distance, although physical distance (in both space and time) can constrain functional distance. To illustrate the difference, imagine a messenger who must travel across a ten kilometer desert, then a two kilometer forest. The message will not get

through if the messenger is attacked by bandits. Bandits are ten times more likely in the forest than in the desert. Thus, the distance through the forest is functionally longer, because there is twice the probability of being attacked by bandits in the forest. Length in edge replacement corresponds to the latter kind of distance. For instance, even if an email crosses several continents, often the most treacherous place it can cross is in the last mile: the receiving computer and server, which can have spam filters, viruses, and other problems. The part of the path that contains the receiving server and computer might be functionally longer, therefore, than the physical distances traversed by the signal when it is bounced off a satellite.

Physical space can still constrain functional space. For instance, order is preserved. If we disable the sending server, then reintroducing the email after the sending server means it will still reach its destination. This is not true if we disable the receiving server. On the other hand, functional space is more abstract than physical space. For instance, there is not a strict Cartesian space in which we must 'make room' for causes. There is no sense in which two causes that are equidistant from another, and in the same direction, will bump into each other. The exact relation between physical and functional space is matter for further research. For instance, it is an open question whether, and if so under what conditions, physically proximate causes are judged to be more likely to share inhibitors. The conservation between physical order and functional order, will suffice for experimental tests in Chapter 4.

Determinism also has theoretical implications for representations of mechanism. Determinism in edge replacement means that variability comes from hidden causes, and that hidden causes are *real:* they have a specific location and extent in space and time, and thus in functional space. The kind of determinism used in edge replacement implies that people should represent mechanisms in more detail when presented with a variable causal relation, than when presented with a causal relation that always works the same way. It also implies that people need to situate these causes in some kind of order, and decide, at any given time, whether each cause is active. This provides a strong prior distribution over the functional forms that should be expected, in contrast to minimality, which allows a broad range of functional forms as long as the graph is minimal. Chapter 5 will describe experimental tests of these implications, along with the mantra that 'variability implies complexity.'

Finally, the branching character of edge replacement implies that mechanisms tend to be shared among causes and effects. For instance, while minimality assumes by default that collateral effect have independent relations to the main cause, edge replacement tends to assume that they share some of the path to the cause. This amounts to a kind of conservation of mechanism – nature is stingy with her causal processes, preferring to use at least part of a mechanism that already exists. We will test this experimentally as well.

If mechanism is a question, then edge replacement is only part of the answer. While these implications are a step forward in understanding human representations of

mechanism, there are nonetheless some important questions that edge replacement leaves unanswered, at least in its current form, and on its own. For instance, force dynamics remains a more complete model of mechanism in the range of causes that deal with simple, physical causation. Integrating force-dynamical models and edge replacement models may be a fruitful area of future research. Another aspect of mechanism is that mechanism knowledge lets us determine what kinds of changes to a causal system will be relevant to the relations that comprise it. Even if you have never painted a computer, or passed a strong electromagnet over one, you know that the paint is less likely to break the computer than the electromagnet. You have learned what categories of actions are relevant. Integrating edge replacement with models of categorization, may help us better understand this type of learning.

Overall, edge replacement articulates a position on mechanism, which would otherwise be vague: When presented with a causal relation, people initially represent it not as a completely abstract relation, but neither do they necessarily fill in every physical detail. Rather, they imagine a functional space, usually containing hidden structure, through which causal power moves deterministically. Edge replacement makes testable predictions about the way that people do this. At least in the cases we have seen so far, edge replacement makes better predictions than minimality. Because minimality corresponds to the mechanism-free account, we can see these tests as good reason to reject a completely mechanism free account. However, showing that minimality is incorrect does not, by itself, show that edge replacement is correct. The

task in the subsequent chapters will be to further develop and experimentally test some of the novel predictions of edge replacement.

A note about complexity

Many scientists see simplicity as a virtue of good theories and models. To many such researchers, edge replacement may seem overly complicated. This section will address such criticisms. It will argue first that edge replacement is not as complicated as it seems at first. It will also argue that whatever complexity edge replacement does have, is justified.

There is a difference between complexity and complication. Both are daunting, but only complexity can be understood by discerning simple underlying rules. Edge replacement is complex, not complicated: It has a few simple ideas, which motivate the more elaborate structures that we observe in the models. There are two core ideas in edge replacement: Causal relations tend to be arranged in branching streams, and causal relations are deterministic. All other complexity (the parameters, the functional forms, branching, volatility) arises naturally from attempts to formalize these ideas so that they are theoretically clear, and make clear predictions. There may be other ways of formalizing these ideas. This thesis does not defend the idea that edge replacement is the simplest possible way of formalizing them. Further work may discover other better or simpler ways of doing so.

Even in its current form, edge replacement is simpler in many ways than alternatives like minimality. For instance, the functional forms are simpler: only

deterministic OR and AND NOT. While the graphs generated by edge replacement often look more complex than minimal graphs, it is worth asking what complexity means in this situation. To achieve minimal graphs, we must offload the complexity into the functional forms of the edges. It is debatable which is more complex; certainly it is not obvious that minimal representations are simpler in all cases.

There is also the fact that minimality is wrong. The evidence shows that people just don't represent causation this way. Simplicity does nothing to recommend a theory, if introducing the simplicity forces the model to make incorrect predictions. We see these incorrect predictions in nonindependence and determinism effects discussed above. In the absence of an alternative, edge replacement may be the simplest of the models that has not yet been proved wrong.

Overall, the analysis is not about which model is complex, because both have complexity. It is about what kind of complexity is the right kind, to model human representations of causal relations. A more appropriate test than perceived complexity, is the number of parameters used, and the firmness of predictions made. If minimal models are simpler than edge replacement, then edge replacement should require more parameters to fit the data, and should not make stronger predictions in so many cases. Edge replacement does more with less, a hallmark of simplicity. It is possible that some of the resistance to, and apparent complexity of, edge replacement, comes from the fact that it is unfamiliar to many researchers. This is to be expected, but it is not a justifiable reason to argue against the model.

Preview: Experiments

As we have seen, edge replacement specifies and elaborates two relatively independent commitments: determinism, and branching streams. Formally, there is no particular reason why these two commitments cannot be treated as independent. For instance, we could have a generative process that created branching streams with probabilistic functional forms, or graphs that use deterministic functional forms that are otherwise minimal.

The previous section argued that these two commitments are theoretically and intuitively coherent, because together, they give us insight into causal mechanisms. We have also seen that the two commitments work together to create a model that fits much human data, from both adults and children. For instance, if the hidden inhibitor in Mayrhofer et al's experiment were captured by the functional form of an edge, rather than by a deterministic inhibitor with real location in the stream, the location of the branch point would not have the same effect, and the model would not fit the data well. The two commitments work well together, but each is a departure from previous models. If the commitments are correct, they should make new predictions, possibly in unexpected areas.

For this reason, the next three chapters will focus on independently testing edge replacement's two commitments in developmental experiments. Another route would have been to focus on a series of experiments that fine-tune the mathematical formalisms of the model, through additional experiments on adults. This is a worthwhile

area for further research, but not the first priority. One reason for this is that work on children provides us with insight about which commitments are the most basic to causal representations.

Another reason is that a model should be productive. Models are formalized theories; as such they are invariably proven wrong or amended in the long run. Along the way, the best models provide us with new insights, and (arguably) most importantly they generate new experimental data, that would never have been collected otherwise. The best such experiments produce data that can remain interesting long after the model has been falsified. Our first priority will be to see if edge replacement can generate new experimental predictions, that may be compelling even to researchers uninterested in the formal details.

The upcoming chapters will focus on experimentally testing edge replacement's two novel commitments independently. The first is that causal representations naturally have a branching character. We will test this by looking for evidence of stream location effects in experiments with preschool-aged children. A second set of experiments will focus on determinism, by testing how children represent variability in causal relations. Finally, a third set will look at how determinism leads to predictions about variability and complexity in causal mechanisms. None of these experiment would have been conducted, had they not been inspired by the edge replacement model. To preview, each set of experiments will produce results that support edge replacement, and cannot be fit by minimal models.

CHAPTER 3: STREAM LOCATION EFFECTS IN PRESCHOOLERS

Edge replacement formalizes a set of commitments, whose implications lead to improved fits (in comparison to minimality) on existing experimental data. One strong example of this is nonindependence effects. If edge replacement's commitments are correct, then they make predictions that go well beyond nonindependence effects. Namely, our most basic representations of cause and effect should be different than the minimal models that are often used to capture them.

This and subsequent chapters will use the term 'basic' to refer to representations that are present early in learning, when little knowledge exists about a given causal system. Because minimality and edge replacement instantiate different prior expectations about the causal structure of the world, it is about basic representations that the two models make the most divergent predictions. One place to find basic representations is in adult novices. This was part of the rationale for using fantastical cover stories, such as Mayrhofer et al's (2010) mind-reading aliens: No one has real-world experience with these causal systems. Chapters 3, 4, and 5 will focus instead on 3- and 4-yearolds in order to test the properties of basic representations, because 3- and 4-year-olds have so little knowledge of the causal structure of the world.

Regardless of whether preschoolers have the earliest representations of cause and effect (for instance, Oakes and Cohen [1990], among others, show evidence suggesting that infants perceive causal relations), preschool-aged children often present us with the earliest examples of causal reasoning experiments that are analogous to

those that can be conducted with adults. A number of experiments suggest that 4-year-old and especially 3-year-old children have simpler representations of causal mechanisms (that is, intermediate causal structures) than adults, and that younger preschoolers have simpler such representations than older preschoolers (Buchanan & Sobel, in press; Bullock, Gelman & Baillargeon, 1982; Sobel, Yoachim, Gopnik, Meltzoff, & Blumenthal, 2007). If even these basic causal representations do not conform to minimality, but do conform to edge replacement, this will provide a strong test of both models.

Chapters 4 and 5 will focus on testing implications of edge replacement's commitment to determinism. This chapter will focus on edge replacement's commitment that causal relations tend to be organized in branching streams. We will test for this by looking for a stream location effect in preschoolers. (Recall that we used a stream location effect to explain Mayrhofer et al's [2010] data in Chapter 3.) We will begin by reviewing the nature and formal motivation for a stream location effect, before describing an experiment that looks for this effect in 3- and 4-year-olds.

Formal Motivation for Stream location

Stream location is best illustrated by beginning with an example. Consider a causal system about which most adults know little: cellular telephones (Rozenblit &Keil, 2002). You and your friend Bob are both trying to reach a mutual friend, Cindy. Despite multiple attempts, neither of you are able to reach her on your phone. Then, you find out that Cindy's cellphone carrier (a different carrier than that used by either you or

Bob) has been experiencing technical difficulties that have prevented you from getting through. You hear news that the difficulties have been resolved, and you try calling Cindy. You can now get through easily. Intuitively, it seems obvious that Bob should now be able to get through easily as well. On the other hand, imagine a different scenario: You and Bob are both having trouble getting through to Cindy, and you discover that your cellphone battery is low on power. You try replacing your battery with a fresh spare, which seems to resolve the problem: You are now able to reach Cindy easily. In this second scenario, it seems unlikely that the change (adding a battery to your phone) would enable Bob to reach Cindy. Even though you understand little about how the cellphone network operates, you can infer the scope of the change -- which causal relations it will affect (Carroll & Cheng, 2009) --from the location of the change. Specifically, in this common effect network, a *late* change (close to the common effect) was more likely to affect multiple relations, than an *early* change (close to individual causes.) By manipulating the location of the change, we changed the scope of the effect that the change had on the causal system. The rest of this chapter will call the result of such a manipulation a *stream location effect.*

One account of stream location effects is that such inferences are due to specific mechanism knowledge. That is, we may not know a great deal about cellphone networks, but we know enough to make the stream location based inference. Such an account supports the hypothesis that stream location effects, where they exist, are due to specific causal structures that are learned domain by domain. This account is

consistent with minimality, because we could have learned the structures from the statistical properties of the domain. Edge replacement is committed to an alternative account: stream location effects arise because of general constraints on representations of causal relations. According to this latter hypothesis, stream location effects should persist in basic representations: situations where familiarity and expertise are minimized. For this reason, Experiments 1-3 use a novel causal environment to test young children's causal reasoning about stream location.

To look at the formal motivation for stream location, we will consider a stripped-down version of the cellphone example, that does not include a detailed cover story. There are two tokens of a cause (A and B) that bring each about an effect C. We see that both A and B are ineffective in producing C, for several trials. We then make a change to the causal system that is either close to C, or close to A. After this change, we see multiple trials in which A is effective in producing C. We are then asked to predict whether B will be effective in producing C on the next trial. For the sake of brevity, this chapter will refer to an affirmative answer here as 'generalizing,' because we generalize the restored efficacy of one relation, to the other relation. Minimality and edge replacement prescribe different answers when asked to generalize in this scenario.

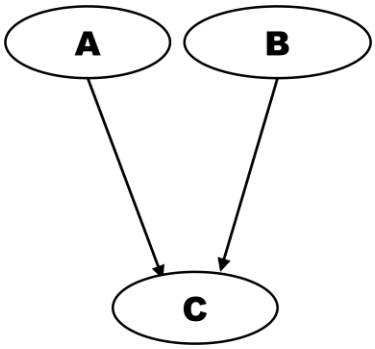

Figure 26: A canonical common effect model.

We can begin by looking at minimality. Assume for the sake of argument that we have no specific knowledge of the mechanism of the system, or of the frequencies of effects given various sets of causes. That is, the scenario is exactly one in which minimality prescribes the simplest, canonical common effect model shown in Figure 26. The stripped-down cover story describes changes that are made to the causal system. Each change must be represented in a CGM using an intervention on a node; interventions on edges are not allowed. Within the constraints presented by the minimal graph, we must intervene directly on either A or C. Both are inconsistent with the data: An intervention on C will sever C's connection with A, but we have seen that after the change, C is effective in producing A. An intervention on A is also inconsistent with the data: According to the cover story, the state of A can still be modified after the change, but after an intervention, it would need to be fixed in order to have an effect on the causal system. Because the strictly minimal graph cannot accommodate the data, we are licensed under minimality to add some complexity. One minimal way of accommodating this is to use the graph shown in Figure 27. In this graph, we have added a node that represents each intervention. Note that there are multiple edges

going in to the node ‘C’, but the functional form is not a simple Noisy-OR: intervening on ‘change near A’ or ‘change near C’ enabled the relation between A and C. There must be some more complex functional form that relates the five nodes.

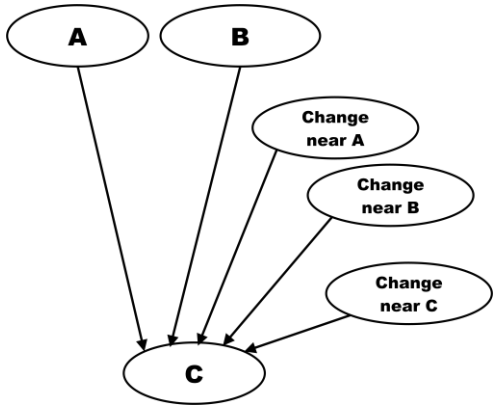

Figure 27: A minimal model that incorporates the interventions from Experiment 1. Note that the functional form is not a noisy-OR.

At this stage, minimality has nothing more to say: It must await more data in order to determine the exact nature of the functional form. For instance, it could be that any intervention is sufficient to enable both relations, or that different combinations of interventions are needed to enable each one. In particular, the minimal graph cannot predict whether the intervention will change the relation between B and C, because it has seen no data on this relation after the intervention. Because minimality insists on minimizing intermediate structure, and allows a range of functional forms, no minimal structure will be able to make a firm commitment about whether we should generalize.

Edge replacement, on the other hand, makes firm predictions here. Figure 28 shows the most likely graph that edge replacement would generate, given the stripped-down scenario with A, B, C, and the interventions. A and B are either generated through

branching, as multiple tokens of the same type of event, or one is generated as a side effect – either process can produce a graph like Figure 28. Because functional space is constrained by physical space, nodes must be arranged in the order shown. For instance, a graph in which 'change to A' and 'change to C' switched places, would be inconsistent with the physical constraints, because C must occur downstream from A. [11]

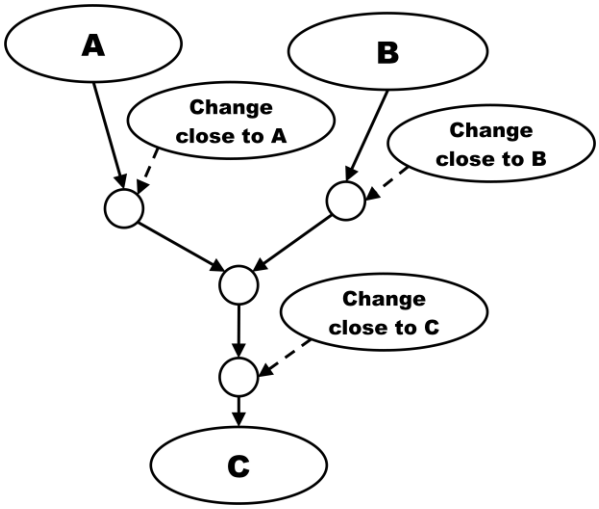

Figure 28: A graph, generated by edge replacement, that fits the data presented to children in Experiment 1.

Because of edge replacement's restricted functional forms, and the rule that inhibitors must be introduced through replacements, the interventions must be generated as inhibitory replacements along the causal stream. Because causal relations are arranged in branching streams, the path from A to C, and the path from B to C, must share some length on which a common inhibitor might fall. It is possible that all three

[11] Note that we are relying here only on the most general and uncontroversial mapping between functional and physical space: functionally prior events must be physically prior. As discussed in Chapter 3, (p. 113) and taken up again in the conclusion, further work can more completely address the mapping between objects in physical space, and edge replacement's representation of events in functional space.

inhibitors, or none of the inhibitors, could fall on this path. In general, the hypothesis space contains more graphs in which 'Change close to C' disables both relations, than graphs in which 'change close to A' disables both relations. This is because the branch point is equally likely to occur anywhere, but C is more likely to occur late in the causal stream. Thus, the branch point is more likely to occur before 'change to C,' than it is to occur before 'change to A.' Edge replacement predicts that interventions on 'change close to C' are most likely to change both relations, while interventions on 'change close to A' are most likely to change just one relation. According to edge replacement, it is more rational to generalize when the change is close to C.

Edge replacement makes the converse prediction in the case of a common cause of two effects. If we reverse all the generative arrows in Figure 28, then we can see that changes close to C (which is now the common cause) will be most likely to change multiple effects. This is related to the way in which edge replacement accounts for the results of Mayrhofer et al's (2010) experiment, in which descriptions of the mechanism changed the degree of nonindependence observed. In the 'sending' condition, edge replacement posits a single, early inhibitor, while in the 'reading' condition, edge replacement posits three separate, late inhibitors. In this case, edge replacement explains Mayrhofer et al's data The experiments below more directly test edge replacement's novel predictions.

Note that edge replacement generates the stream location effect even with this stripped-down cover story that uses blank events and limited data. The stream location

prediction arises from the structural constraints present in edge replacement. Minimality, on the other hand, makes no firm predictions. Minimality predicts that in situations of limited knowledge, people should not show stream location effects. The more limited the knowledge, the more divergent the predictions of the two models become. For this reason, Experiments 1-present preschool-aged children with a scenario that follows the pattern described in this section, employing progressively more unfamiliar mechanisms over three experiments. Experiment 1 uses batteries, a mechanism about which both 3- and 4-year-olds can reason appropriately about in this experimental paradigm (Buchanan & Sobel, in press). In Experiment 2, adding a cover to the lights apparently restores the efficacy of a causal relation; this is a more unfamiliar intervention. Finally, in Experiment 3, removing a battery apparently restores the efficacy of a causal relation – this is the most unfamiliar mechanism with which we present children, because it is contrary to their existing knowledge. Edge replacement predicts that children should show a stream location effect, even as the mechanisms became more unfamiliar.

Experiment 1

The first experiment presented children with a novel causal system that conformed to the cover stories described above (i.e. Figure 28). Children were shown a common effect structure in which both causal relations failed initially: Neither cause was successful in producing the effect. The experimenter then made a change (adding batteries) that enabled one of the causal relations. After the change children saw that one of the

causes was now effective in producing the effect. Children were then asked to predict the efficacy of the other cause. According to edge replacement, experimentally manipulating the location of this change in the causal mechanism should affect responses: Children who observe a change close to the common effect should be more likely to generalize, than children who observed a change closer to the each individual cause.

Methods

Participants

Thirty-two preschoolers (22 boys, 10 girls, M = 46.32 months, Range = 36-59 months) were recruited from birth records, a local preschool, and the Providence Children's Museum. Three additional children were tested, but were excluded due to experimenter error or equipment failure. Children were randomly assigned to either the early (n = 16) or late (n = 16) condition. There were an equal number of 3- and 4-year-olds in each condition. The racial and ethnic breakdown of the sample was as follows: 29 children were Caucasian, 2 were African-American, and one child was Hispanic/Latino. No information about socio-economic status was collected.

Materials

Materials consistent of two sets of commercially available closet lights, modified for the experiment. One set (the 'cause lights') consisted of eight lights, each 10 cm in diameter, with a button that illuminated when pressed (See Figure 29) The underside of

each light held a battery compartment enclosed by a removable cover that could be left on or off (leaving the batteries visible). The casing of each light was painted a different color.

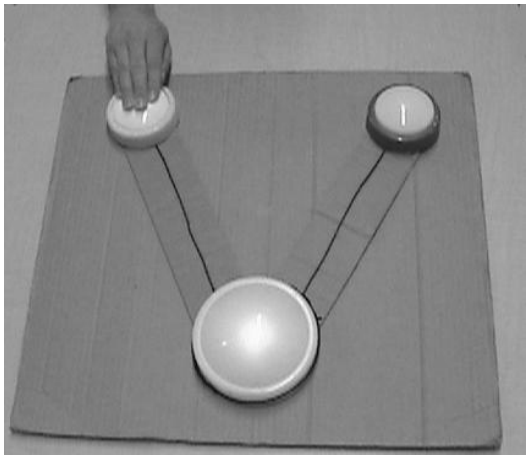

Figure 29: The lights used in Experiments 1-4, shown from the child's point of view.

The other set comprised four more lights, the 'effect lights.' These were larger, measuring 14 cm in diameter. One effect light was modified such that it only activated via a remote control. The experimenter practiced an effect in which the light appeared to activate when a participant made an action (such as pressing the effect light, or pressing another light). This effect was convincing to both children and adults, who often reported surprise when the remote was revealed. The remote allowed the experimental to control which actions, if any, would appear to cause the light to activate.

The lights were mounted together three at a time on a piece of cardboard, apparently connected by black wires, as shown in Figure 29. Each effect light had a pipe cleaner wrapped around its casing so that children could tell them apart, but the apparent 'identity' of the light could be switched surreptitiously. All four lights were

shown together at the outset of the experiment, but the one light that was activated with the remote, was the only one actually used in the test phase. Because of the deception, children believed that they were being shown the four lights in succession.

Procedure

The experimenter familiarized the children with the environment, saying: “I have some of these lights. When you push on them, they light up. See [pushes on effect light, and it illuminates]. Here, you try.” All children pushed the effect light, which illuminated when depressed. “Sometimes, when I push on the little lights, they make the big light go. Watch.” The experimenter then pushed each cause light, which each appeared to immediately cause the large light to illuminate. He then pointed to each of the cause lights in turn and asked, “Does this one make the big light go?” Most children (26 out of 32) correctly answered “yes” to both of these questions. The remaining children responded correctly after one instance of corrective feedback. Excluding children who required feedback on this or any other training question did not change the statistical significance of any results reported below.

The experimenter then removed the three initial lights, and arranged three apparently new lights in the same configuration. This began the first of three test trials. In the ‘late’ condition, the effect light in each test trial was missing a battery. In the ‘early’ condition, one of the cause lights in each test trial was missing a battery. Lights without batteries (including the cause lights) did not illuminate when pressed. Battery

covers were left off, so that children could see the absence of batteries when the lights were flipped over.

For each test trial, the experimenter began by depressing the cause lights, demonstrating that they failed to activate the effect light. Note that in the early condition, the cause lights did not contain batteries, and so did not illuminate when depressed. In the late condition, cause lights did illuminate when depressed. Due to the nature of the lights (i.e. they cannot operate without batteries) this was an unavoidable difference between conditions. The child was asked to verify whether each cause light made the effect light activate. Most children (26 out of 32) correctly answered "no" to these questions on all three trials. Five children responded correctly after one round of feedback, and one child required two rounds. Excluding these children did not change the significance of reported results. Note that at this point, all children had correctly answered "no" to two questions with feedback, and "yes" to two questions with feedback. This means that children were not coached on a strategy that would always allow them to answer the test questions correctly.

The experimenter then made a change to the causal system, depending on the condition. In the early condition, he flipped over one of the cause lights, and said: "Look, this light has room for a battery, but there's no battery there. Let's put a battery in." He inserted a battery, and flipped the cause light back over. The battery was inserted backwards, so that the light would not begin illuminating when pressed; this would have allowed children to deduce that the other cause light, which did not illuminate when

pressed, would be ineffective. In the late condition, the experimenter inserted a battery into the effect light (as opposed to one of the cause lights). In the late condition, of course, the batter was correctly inserted.

The experimenter then said: “Now let’s see what they do.” He pressed one of the cause lights (in the early condition, always the one that had been changed) and it apparently caused the effect light to activate. The experimenter asked, pointing to the light he had just pressed: “Does this one make the big light go now?” All children correctly answered “yes” to this control question. He then asked the test question about the other light: “What about this one? Will this one make the big light go now?” Responses to this test question were recorded and analyzed by a research assistant who was blind to the experimental hypotheses. A second research assistant coded a random sample of 25% of the data; inter-rater agreement was 100%. After the test question, the experimenter moved to the next test phase without giving feedback or showing the actual efficacy of the lights: he removed all three lights, and brought out three apparently different lights, repeating the procedure. This was done twice, for a total of three yes/no questions from each child.

Results

Responses to the test questions are shown in Table 1. Preliminary and chance analyses considered the proportion of children that gave correct answers to all three test questions. This was defined as three “yes” answers in the late condition, or three “no” answers in the early condition. Twenty-nine out of the 32 children generated this

pattern of response. There was no significant difference in this distribution between genders, Fisher's Exact test, p = 1.00, between age groups, Fisher's Exact test, p = 1.00, or between conditions, Fisher's Exact test, p = 0.23. These data were compared to the proportion of all correct answers that would be expected if children guessed on each trial (0.125), if they randomly chose an initial response and then perseverated (0.50), or if they simply said 'yes' to all questions (0.50). The data were significantly different in each case, Binomial tests, all p-values < 0.01.

The next step was to compare the number of "yes" responses between the two conditions. The overall distribution of "yes" responses differed between the conditions, $\chi^2(3, N = 32) = 32.00, p < 0.01, 0.71$. Because of low expected counts,this analysis was supplemented by collapsing children into two categories: Those who answered "yes" to every test question, and those who showed any other pattern of responding. In the late condition, children generated this pattern of response 100% of the time, while children in the early condition generated this pattern of response 0% of the time, a significant difference, Fisher's Exact test, $p < 0.01$. Similar results were found when considering the proportion of children who said "no" to every test question as opposed to those who showed any other pattern of responding: Children in the early condition generated this pattern of response 81% of the time while children in the late condition generated this response 0% of the time, also a significant difference, Fisher's Exact test, $p < 0.01$.

Table 1: Distribution of responses in Experiments 1-2

	Number of "yes" responses on test questions			
	0	1	2	3
Experiment 1 (Familiar Mechanism)				
Early (n = 16)	13	3	0	0
Late (n = 16)	0	0	0	16
Experiment 2 (Unfamiliar Mechanism)				
Early (n = 16)	11	3	2	0
Late (n = 16)	1	0	1	14

Discussion: Experiment 1

The data show floor and ceiling results: The majority of children in the early condition said “no” to all test questions; the majority of children in the late condition said “yes” to all the test questions. Under edge replacement, this pattern of data occurred because children inferred that a change closer to the common effect would affect both relations, while a change closer to one of the individual causes would affect only that specific causal relation. Other accounts, including minimality, could explain these data by positing that children had learned about stream location effects due to extensive experience with batteries. Previous research (e.g., Buchanan & Sobel, in press; Gottfried & Gelman, 2005) suggests that children as young as 3 understand batteries to make appropriate inferences about relevant and irrelevant modifications to a causal system when batteries are involved. In order to test whether causal stream location is a general structural principle in children’s reasoning, we need to look for a similar effect in an unfamiliar causal mechanism, where such experience would not be present.

Experiment 2

The goal of Experiment 2 was to test whether children would still reason in accord with stream location when the change was unfamiliar. It used a similar environment to Experiment 1, but no lights were missing batteries – instead, some were missing a battery cover. Adding a cover to the lights apparently restored their causal efficacy. If

children recognize stream location as a general structural principle, this procedure should replicate the results from Experiment 1.

Method

Participants and Design

Thirty-two preschoolers (8 girls, 24 boys, M = 46.44 months, Range= 36-57 months) were recruited from birth records, a local preschool, and the Providence Children's Museum. One additional child was tested, but was excluded due to experimenter error. Included participants were randomly assigned to either the early (n = 16) or late (n = 16) condition, with an equal number of 3- and 4-year-olds in each condition. The racial and ethnic breakdown of the sample was as follows: 28 children were Caucasian, 1 was African-American, 1 was Asian, 2 were Hispanic/Latino, and 2 were of mixed race. No information about socio-economic status was collected.

Materials and Procedure

The materials and procedure were the same as in Experiment 1 except for two changes. The first change was that all of the lights had each battery slot filled (the battery covers were still initially left off). Second, the intervention that apparently changed the efficacy of the relation was not adding batteries, but adding a cover to the opening that housed the batteries. The experimenter said: "Look, this one is missing a cover. Let's put a cover on here." Otherwise, the procedure was the same as in Experiment 1: Children were initially shown a set in which two cause lights were effective, and were asked to verify

the efficacy of the lights. Most children (30 out of 32) answered correctly; the remaining two children answered correctly after one round of feedback. Children then participated in three test trials. In each test trial, both cause lights were ineffective, and children were asked to verify this. Most children (22 out of 32) correctly answered "no" to both of these questions on all three trials. Six children required one round of feedback, two children required two rounds of feedback, and two children required three rounds or more. Excluding all children who required any feedback does not change the statistical significance of the results reported below.

The experimenter then showed that adding a cover (either to the cause or effect light, depending on the condition) appeared to enable one relation, and asked children to verify that the light had become effective (no children failed to answer this question correctly). The experimenter then asked about the efficacy of the other relation. Responses to this test question were recorded and analyzed by a research assistant who was blind to the hypothesis. A second rater coded a random sample of 25% of the data; inter-rater agreement was 96%. The experimenter resolved the disagreements by looking at the videos. Finally, to ensure that the experiment did not negatively affect children's causal knowledge, all children were briefed on the deception involved in the experiment: Children were shown the remote and allowed them to play with it at the end of the procedure.

Results

Responses are shown in Table 1. The data were analyzed in the same manner as in Experiment 1. Twenty-five out of 32 children answered correctly to all three test questions. There was no significant difference in this distribution between genders, Fisher's Exact test, $p = 0.63$, between age groups, Fisher's Exact test, $p = 1.00$, or between conditions, Fisher's Exact test, $p = 0.39$. This pattern of responding was compared to the proportion of all correct answers that would be expected if children guessed on each trial (0.125) or if they randomly chose an initial response and then perseverated (0.50), or if they showed a yes bias (0.50) – the data were significantly different in each case, Binomial tests, all p-values < 0.01.

The next step was to compare the number of "yes" responses between the two conditions. The distribution of the number of "yes" responses differed between the early and late conditions, $\chi 2(3, N = 32) = 25.67, p < 0.01, \phi = .51$. Because of low expected counts, this analysis was supplemented by collapsing the data as in Experiment 1 – comparing the number of children in each condition who responded "yes" to all three test questions. In the early condition, children generated this pattern of results 0% of the time; in the late condition, children generated this pattern 88% of the time, a significant difference, Fisher's Exact test, $p < 0.01$. A similar finding was obtained when by analyzing whether children said "no" to every test question (69% in the early condition vs. 6% in the late condition), Fisher's Exact test, p = 0.01.

Discussion: Experiment 2

As in Experiment 1, children in Experiment 2 were more likely to generalize in the late condition than in the early condition. These data further support edge replacement's predictions about basic representations, by showing that children still show stream location effects when the change is one that does not normally affect efficacy.

One potential objection to the relevance of these results is that children may have experience with battery covers. For instance, they may have seen an adult remove the cover from a toy, exchange the batteries for fresh ones, and then replace the cover. Children might have developed a mistaken association between changing the cover and changing efficacy. To address this issue, in Experiment 3 asked preschoolers to reason about a change that genuinely conflicts with their existing mechanism knowledge: removing batteries apparently enabled a causal relation. This is certainly a scenario with which children could not have had experience; a stream location effect here would be strong support for the idea that basic representations show stream location effects.

Experiment 3

Methods

Participants

Thirty-two preschoolers (17 girls, 15 boys, M = 48.00 months, Range= 36-60 months) were recruited from birth records and a local children's museum. Three additional children were tested, but were excluded because they failed controls. The racial and

ethnic breakdown was as follows: 25 children were white, 2 were African-American, 2 were native Hawaiian or other Pacific islander, and 2 were of mixed or other races. No information about socio-economic status was collected.

Materials

For this experiment, the lights were modified such that they had a false bottom. The light was actually powered by lights hidden deep inside the light, unbeknownst to the child. The visible battery compartment could contain a second set of batteries that were actually unrelated to the activation of the light. The false bottom made the lights about 2 cm taller, but they were otherwise similar to those in Experiments 1-2. The experiment used two effect lights, and four cause lights. As before, the effect lights were individuated by a pipe cleaner placed around their body. Each cause light was painted a different color.

The false bottoms allowed Experiment 3 to address a potential concern with previous experiments. In Experiments 1-2, the cause lights illuminated in the late but not the early condition. Although there is no obvious reason why this would bias the data, Experiment 3 presented the opportunity to make the early and late conditions as similar as possible. In both conditions of Experiment 3, each light always lit up when pressed, regardless of whether it had batteries in the visible compartment. This meant that there was no difference between conditions, in whether the cause lights illuminated when pressed.

Procedure

Because of the robust effects in previous experiments, children received only a single test trial. The experimenter first familiarized children with the push lights environment as before, this time using the modified lights. During the familiarization phase, the experimenter showed children two cause lights that activated an effect light, and asked training questions with feedback. Most of the children (28 out of 32) required no feedback in this experiment; the remaining four children required one round of feedback before they answered correctly. All reported significant results remain statistically significant when these four children are excluded.

For the test trial, the experimenter showed children two new cause lights that failed to activate a new effect light, and asked children to verify their efficacy. Most children (27 out of 32) correctly answered "no"; four children required one round of feedback before answering "no," while one child required two rounds. Two children (both 3-year-olds) were excluded due to a failure to generate the correct response within four trials of feedback.

The experimenter then said: "Let's look underneath this light. Look, this light has batteries. Let's take the batteries out." For the children randomly assigned to the early condition, the experimenter removed the batteries from one of the cause lights, showed the empty compartment to the child, and replaced the light in its original location. For the remaining children in the late condition, he did the same thing to the effect light. The experimenter then demonstrated that this enabled one of the relations – that

specific cause light (in the early condition) or one of the cause lights (in the late condition) now made the effect light activate. The experimenter asked children to verify this change in efficacy. One child's data was replaced because she failed to answer this correctly (no feedback was provided).

Children were asked about the efficacy of the other causal relation, and their answer was recorded. Responses were coded by research assistant who was blind to the hypothesis. A second rater coded a random sample of 25% of the data; inter-rater reliability was 100%. To avoid biasing children's causal reasoning with misleading data, all children were briefed on the deception: The experimenter showed them that the lights had more batteries deeper inside, and allowed them to play with the remote.

Results

Responses to the test question are shown in Table 2. No effects of gender or age group were found on the proportion of children responding "yes" to this question, Fisher's Exact test, all p-values > 0.32. One out of 16 children (6%) said this light would be effective in the early condition, compared with twelve out of 16 children (75%) who made this response in the late condition, Fisher's Exact test, $p < 0.01$. Responses in both conditions were significantly different from chance levels (50%), Binomial test, both p-values < 0.01.

In order to compare the data between all three experiments, these analyses were repeated on the first trials of Experiments 1 and 2. Results from the first trial of Experiments 1-2 are also shown in Table 2. In all three experiments, significantly more

Table 2

Responses to Test Question in Experiment 3 Compared to First Test Trials in Experiments 1-2

	Frequency of "yes" responses on first test trial	
	Early condition	Late condition
Experiment 1 (Adding a battery)	1 out of 16	16 out of 16
Experiment 2 (Adding a cover)	4 out of 16	14 out of 16
Experiment 3 (Removing a Battery)	1 out of 16	12 out of 16

children answered “no” in the early condition and “yes” in the late condition on the first trial than would be predicted by chance, Binomial tests, all p-values < 0.01. In both Experiments 1-2, children were significantly more likely to answer “no” on the first trial in the early than in the late condition, Fisher’s Exact tests, all p-values < 0.01.

Discussion: Experiment 3

In Experiment 3, children were again significantly more likely to generalize when the change was closer to the common effect (i.e., late in the causal stream) than when it was closer to an individual cause (i.e., early in the causal stream). This was true even though the results of that change conflicted with their existing mechanism knowledge – if children saw a single example in which removing a battery enabled a causal relation, they could incorporate the location of that change into their inference about the efficacy of that manipulation. Because children could not have observed such a change having such an effect outside of the experiment (i.e. removing batteries does not enable causal relations between electronic toys), they must have made the inference using more structural knowledge about causation in general, as edge replacement suggests.

Discussion of Experiments 1-3

Three experiments found strong stream location effects in both 3- and 4-year-olds, whether the mechanism was consistent with children’s knowledge (Experiment 1), novel, (Experiment 2) or even contrary to their existing mechanism knowledge

(Experiment 3). In each case, the location of the change determined whether children would generalize, as predicted by edge replacement: Children were more likely to predict that the BC relation would be restored, if the change that restored AC was close to C, than if it was close to A.

Taken together, these data show that by age three, children have a relatively general piece of causal knowledge that gives rise to stream location effects. The knowledge cannot be mechanism-specific, because Experiments 2 and 3 replicated the effect with novel mechanisms. It remains to be experimentally shown exactly how general this knowledge is. For instance, it is possible (but unlikely) that children will fail to show stream location effects in other domains, or other causal environments, than the lights from this experiment. This experiment has yet to be conducted with a common cause structure; presumably children will reason in accord with stream location here as well. More work is needed to fully establish stream location as a general principle of preschooler's causal reasoning.

These findings are sufficient, however, to provide an initial test of edge replacement as compared to minimality. As shown at the beginning of the chapter, minimality cannot fit these findings: The minimal structures associated with the relations shown to children, are unable to make firm predictions about the BC relation, until they have seen data about the efficacy of the BC relation. Though they seem straightforward, and even obvious, children's answers are not predicted by minimality. Edge replacement, however, does predict children's responses, as shown at the

beginning of the chapter. To recap: Edge replacement holds that children infer that an intervention that is physically closer to one of the cause lights than the effect light, will be functionally closer to the nodes that represent events related cause light, than to the nodes that represent events related to the effect light. Because of the branching character of causal structures created by edge replacement, these interventions are most likely to fall on the part of the path not shared by the other relation. Therefore, the intervention will likely restore only one relation. The converse prediction also arises from edge replacement: interventions close to the effect light will likely restore both relations. These predictions arise naturally from edge replacement, even when the model is asked to reason about totally novel causal scenarios, as children were asked to do here.

Another aspect of Experiments 1-3 is worth noting: They show an indirect test of children's commitment to causal determinism. Each experiment showed children a causal relation that changed, from consistent failure, to consistent success. In principle, this could happen randomly, just as a coin can produce a long sequence of tails followed by a long sequence of heads. The data do not support the idea that children represented relation as random: Randomness would imply that the intervention was irrelevant, but children generalized differently depending on the location of the intervention. This is particularly salient in Experiment 3, where the intervention was counter to children's mechanism knowledge (removing a battery): Children were apparently so reluctant to accept a random or unexplained change in efficacy, that they were willing to attribute

the change to an intervention they knew should have been ineffective. The next experiments will focus on testing children's determinism more directly.

CHAPTER 4: DETERMINISM AND HIDDEN INHIBITIORS

Edge replacement captures a set of commitments about the character of early representations of cause and effect. Chapter 4 focused on the stream-like character that edge replacement predicts should exist in basic representations: Children should expect, readily and by default, that collateral effects are generated through branching. Experimental tests of predictions arising from this commitment, supported edge replacement over minimality. This chapter will investigate a separate commitment that edge replacement also instantiates: That children should expect, readily and by default, that direct causal relations are deterministic. In particular, children should readily posit hidden inhibitors, not intrinsic randomness, to explain variability in causal relations. We will see that this commitment leads edge replacement to make different predictions than minimality in experimentally testable ways. As before, we will test these predictions on preschoolers, because these early representations are likely the most basic, leading the two models to make the most divergent predictions.

Recall from Chapter 3 that edge replacement does not allow probabilistic functional forms. Rather, all edges must have a deterministic functional form. In order to capture an apparently probabilistic relation (for instance, the effect occurs 7 times out of 10 given the cause), edge replacement must posit a hidden inhibitor (for instance, one that has a probability of firing of 0.3, or multiple inhibitors that achieve the same effect). Recall also from Chapter 3 that edge replacement says that these inhibitors are *real:* they have a concrete existence in space and time. The temporal component of

edge replacement describes how the activation of hidden inhibitors should change over time. We introduced and formalized the concept of *volatility* to capture this commitment: In systems with low volatility, hidden inhibitors persist in time: A given inhibitor is more likely to be active, given that it was active on the previous trial. In the limit case in which volatility is infinite, edge replacement recovers complete randomness, in which trials are independent. Assuming that volatility is finite, edge replacement prescribes that children should posit real, persistent inhibitors, to explain variable causal relations.

This chapter will compare edge replacement's predictions to those of minimality. Consider the example above, in which the effect occurs 7 times out of 10, given the cause. With unrestricted functional forms, we can accommodate the relation using a single edge, with a functional form that captures a probabilistic relation. Because such a graph uses fewer edges, minimality prescribes that we should use it by default in basic representations. We will assume that minimality also uses by default the simplest functional form available: Assigning a probability to the functional based on the proportion of successes to total trials. In this case, we would assign a probability of 0.7 to the edge. For this reason, the rest of the chapter will refer to minimality as a 'proportional' model. The arguments in this chapter apply to any model that takes the approach of using such a proportion at its core. Note that minimality does not capture the persistence of inhibitors, because it does not posit inhibitors. This will be the source of divergent predictions, between edge replacement and minimality.

One example of such divergent predictions occurs we must reason about a novel intervention. For instance, imagine you own a new car that has always started reliably. On Monday, your friend, an amateur mechanic, opens the hood and makes some 'improvements.' On Tuesday, your car fails to start. Without further repairs, it seems unlikely your car would start on Wednesday. On the other hand, imagine if the car was old junker, that normally only started half the time: it had failed on Thursday and Saturday, but started on Friday and Sunday. As above, your mechanic friend fiddles under the hood on Monday, and on Tuesday it fails to start. In this case, it seems less likely that the car will fail again on Wednesday. The fact that the car sometimes failed before, then recovered, makes you more confident that it will recover again.

In this example, note that adding failures to a sequence seems intuitively to increase, not decrease, the probability of a subsequent success. Upon reflection, this seems to contradict the idea that the probability of success is based on the proportion of successes so far. We will verify mathematically below that such an intuition in the car example cannot be captured by proportional models, but can be captured by edge replacement. In order to first establish exactly the phenomena we wish to model, this chapter will begin by describing an experiment that explicitly tests whether children reason in this way. Subsequent sections will model the results.

Experiment 4 randomly assigned children to one of two conditions: In the *solid* condition, children were presented with a sequence of successes, followed by a change, followed by a failure. They were then asked to predict whether a success was likely on

the next trial. The *sporadic* condition was identical, except that the sequence contained some failures before the change, arranged to suggest great variability. This chapter will begin by describing the methods and results of this experiment. We will then turn to examining edge replacement's predictions in detail. Edge replacement predicts that children could be more likely to predict success in the sporadic condition, despite its lower proportion of previous successes.

Experiment 4

Methods

Participants

Participants included 32 three-year-olds (M=41.50 months, Range: 36-47 months, 12 girls), and 32 four-year olds (M=53.3 months, Range: 48-59 months, 18 girls). Four additional children (all three-year-olds, two in each condition) were tested, but excluded because they failed controls (described below). Half the included children in each age group were randomly assigned to the solid condition, while the other half were assigned to the sporadic condition. About half of the children were tested in a laboratory at Brown University, while the remaining children were tested in a quiet room at the Providence Children's Museum.

Materials

This experiment used the same set of lights and wires used in Experiments 1 and 2, (shown in Figure 29) with one modification: The effect light had a dummy switch installed inside of the battery compartment that was not connected to the wiring of the light. Experiment 4 used two 'effect' lights, one controlled with a remote, and four 'cause' lights. Using the remote, the experimenter could control when and if pressing on the cause light appeared to cause the effect light to illuminate (a 'success') or did not cause the effect light to illuminate (a 'failure').

Procedure

Training Phase

The experimenter first brought out the effect light, and demonstrated that pressing on the effect light caused it to activate. He allowed the children to press the light as well, which illuminated when they pressed it. (This was achieved using the remote.) The experimenter then held his hand over the light, and asked: "I'm going to press on this light again. Will it light up?" If children answered correctly, saying "yes," the experimenter proceeded to the next part of the experiment. If they answered incorrectly, gave a nonverbal response, or did not answer, the experimenter gave children corrective feedback, pressed the light again, and asked the question again. If children needed more than three instances of corrective feedback, their data were excluded from the experiment for failing the controls. (This happened four times, each

time with a 3-year-old.) The experimenter repeated the training phase with the decoy light, which did not illuminate. In this case, the correct answer was “no.” Note that at this stage, all children who passed controls had correctly answered “yes” to one question, and “no” to another. This ensured that children using the strategy of answering “yes” to every question, were excluded.

Test Phase: Solid Condition

To begin the test phase, the experimenter brought out three apparently new lights, and arranged them as in Figure 29. The experimenter said: “Sometimes, when you push on the little lights, they make the big light go. Let's see what they do.” He first pressed three times on the light to the child's right, which apparently caused the effect light to illuminate each time. He then pressed three times on the left light, with the same result. Finally, the experimenter pressed on the right light three more times, again causing the effect light to illuminate each time. He then picked up the effect light, and said “I'm just going to do something to the light.” He held the light vertically, such that the underside was hidden to the child, opened the back cover, and flipped a switch inside of the battery compartment with an audible click. (The switch had no real electronic effect or purpose.) The experimenter then replaced the light on the cardboard, and pressed on the right light. The right light illuminated, but the effect light did not. The experimenter then asked, pointing to the right light: “I'm going to push on this one again. What do you think: If I push on this one again, will it make the big light go?” He then repeated the question, referring to the left light. Children's answers to this test question were

recorded. The experimenter then repeated the whole test phase with an apparently new set of three lights, starting with the light on the child's left. This yielded a total of four yes/no questions from each child. The colors of the lights were randomly counterbalanced, as was the starting location of the light referred to here for simplicity as the 'right' light. That is, about half the children saw the procedure start with the light on their left, and about half the children saw the procedure start with the light on their right.

Test Phase: Sporadic Condition

Children in the sporadic condition were given the same procedure, except that the effect light sometimes failed to illuminate when the cause lights were pressed, even before the experimenter flipped the switch. We can see the solid condition as a series of nine successes, followed by an intervention, followed by a failure, followed by a trial for which a prediction was requested. This will be represented as: [1 1 1 1 1 1 1 1 1 | 0 ?]. The sporadic condition always followed the pattern [1 0 1 1 1 0 0 1 1 |0 ?]. That is, the right light succeeded, failed, then succeeded, the left light succeeded, then succeeded, then failed, and finally the right light failed, then succeeded, then succeeded. The sporadic sequence was constructed to appear random, containing an appropriate number of spontaneous alternations (4/8). It also terminated with two successes before the intervention. Note also that in the sporadic condition the total proportion of successes was the same for each light (2/3).

Results

A research assistant assigned each child a score, counting the total number of times that the child answered "yes" to the test questions (range: 0 to 4). A second research assistant coded a random sample of 25% of the data – inter-rater agreement was 100%. Means are shown in Figure 30.

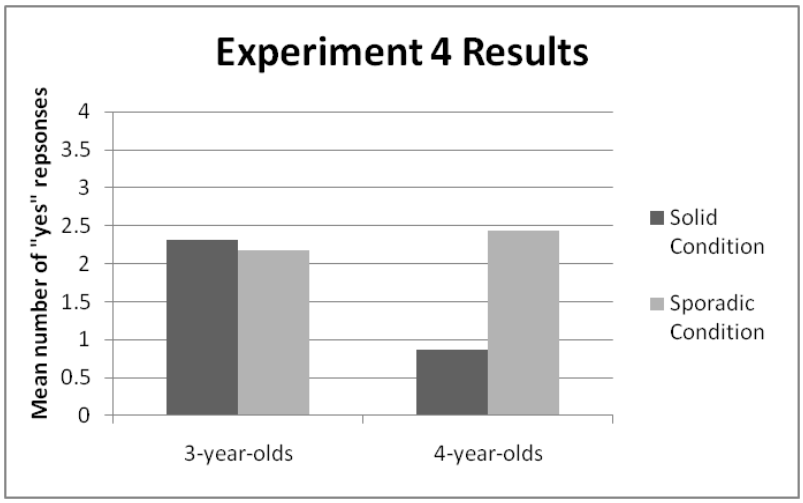

Figure 30: Results from Experiment 4.

Preliminary analyses revealed no effects of gender, or counterbalanced factors such as the side of the first demonstrated light. The data were subjected to a two-way (age × condition) ANOVA. This revealed a main effect of condition, ($F(1,60) = 5.71, p = 0.020$), a marginally significant but statistically inconclusive effect of age group ($F(1,60) = 2.83, p = 0.098$), and a significant interaction between age group and condition ($F(1,60) = 4.148, p = 0.046$). Because of the interaction, simple effects of condition were investigated within each age group. There was a significant effect of condition for 4-year-olds, ($t(30) = 3.30, p < 0.01$), but not for 3-year-olds ($t(30) =$

$0.239, p = 0.81$). This difference arises from the fact that 4-year-olds were more likely to say “yes” in the sporadic, than in the solid condition. Simple effects of age group were also examined, within each condition. There was an effect of age group within the solid condition ($t(30) = 2.87, p < 0.01$), but not within the sporadic condition ($t(30) = -0.23, p = 0.82$).

Subsequent analyses verified these results by looking at individual questions. Children were asked four questions: about the light they had seen fail after the intervention on the first trial (‘demo-1’), the light whose efficacy they had not seen after the intervention (‘other-1’) and the corresponding lights on the second trial (‘demo-2’ and ‘other-2’). Within each age group, the pattern of yes/no responding on each question between conditions, were subjected to a two-sided Fisher's exact test. Three-year-olds showed no significantly different pattern of responding between conditions, on any of the four test questions (all *p*-values greater than 0.05). Four-year-olds, however, showed significantly more “yes” answers in the sporadic than the solid condition, on three out of the four questions: demo-1 ($p = 0.02$), demo-2 ($p = 0.015$), and other-2 ($p = 0.011$). They did not show a significant difference on other-1 ($p = 1.00$). This may be because some children felt the need to balance their initial “no” response, with a “yes” response. Importantly, this did not seem to happen on the second trial (when presumably it became obvious that most answers were “no”). Most importantly, note that 4-year-old children showed a significant difference between conditions on the very first question asked.

Fisher's exact test was also used to further investigate the developmental difference found in the solid condition. Individual questions showed a significant difference between age groups in the solid condition, using Fisher's exact test, for demo-1 ($p = 0.037$), and demo-2 ($p = 0.037$), but only marginally for other-2 ($p = 0.066$) and not for other-1 ($p = 0.45$). No developmental differences were found in the sporadic condition.

Chance analyses were in accord with these results. Chance was defined as a 50 percent probability of saying "yes." Binomial tests on the pattern of responding on each question, within each age group and condition, did not reveal any questions on which 3-year-olds responded significantly differently from chance. The same was true of 4-year-olds in the sporadic condition. In the solid condition 4-year-olds were significantly more likely to answer "no" than chance would predict, on demo-1 ($p < 0.01$), demo-2 ($p < 0.01$), and other-2 ($p = 0.021$), but not on other-1 ($p = 0.21$).

Overall, 4-year-olds showed a clear tendency to respond "yes" more often in the sporadic condition, than in the solid condition. Three-year-olds did not show this tendency -- they were not significantly different from chance in either condition. In the solid condition, there was a significant developmental difference between 3- and 4-year-olds.

Model Description and Results

This section will describe the predictions made by minimality and edge replacement with respect to Experiment 4. To preview, each class of model will make internally

unanimous predictions, but the predictions will be different between minimality and edge replacement.

Some assumptions will be common to both models. The first is that the probability of the effect in the absence of the cause is zero. This is natural given that children never see the effect light illuminate without pressing the cause lights. Second, both models will assume for simplicity that children should answer "yes" to the test question if the probability of the effect occurring is greater than some threshold. Preschoolers like to say "yes" (e.g., Heather Fritzley & Lee, 2003), so it is plausible that this threshold is lower than 50 per cent. This analysis will assume that higher probabilities generally correspond to more frequent "yes" responses. Because other factors (such as guessing) may go into children's responses, this analysis will not treat chance-like responding as inconsistent with any model. For instance, adding noise to any model can produce chance-like responding. Because of the thresholding assumption, the models should not be expected to predict exactly the proportion of children that say "yes" – this would only occur if children guessed with a probability proportional to the model's predicted judgment. This is an unwarranted assumption, especially because we have a small sample. We will focus instead on qualitative predictions: On which condition, if any, should children be more likely to say "yes"?

Minimality

We will begin by considering minimality's predictions. Children are shown two causes and an effect in a novel causal system, so minimality prescribes we begin by using the

canonical common effect model shown in Figure 31a. We could optionally include the switch, while still remaining minimal in a looser sense, as in Figure 31b -- we will consider this case separately below. First, we will focus on Figure 31a. Note that no hidden inhibitors are represented in this model, and children are never given statistical evidence (such as a correlation) to suggest that a hidden inhibitor exists. For this reason, minimality is committed to representing the relation between each cause light, and the effect light, as a single edge. In both conditions, variability exists in at least one causal relation: a cause fails at least once to produce the effect. Because no more complex structure is motivated, minimality must use the functional form of the relation, to accommodate the variability. In order to fit the data, minimality faces a challenge: Find a functional form which, if used in both conditions, will predict high probability judgments in the sporadic condition, and low probability judgments in the solid condition.

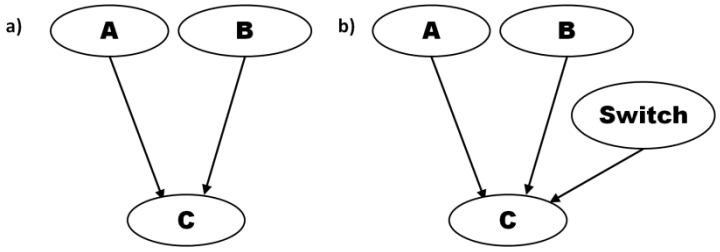

Figure 31: Two minimal models, that could capture the causal system shown to children in Experiment 4.

The simplest functional form for minimality to use is a function of the proportion of successes to failures. For instance, in the solid condition, there were nine successes and one failure, so this approach might predict that the probability of success was 0.9. In the sporadic condition, the proportion was lower: six out of 10, so 0.6. These

predictions are shown in Figure 32: they do not fit the data even qualitatively. In order to fit the data better, we might adjust the function by assigning a strong prior distribution to the functional form. Because this prior would have to be the same between conditions, it will not help minimality fit the data: The proportion is higher in the solid than the sporadic condition, so the best that a strong prior would be able to do is overwhelm these proportions, and predict the same judgment between conditions.

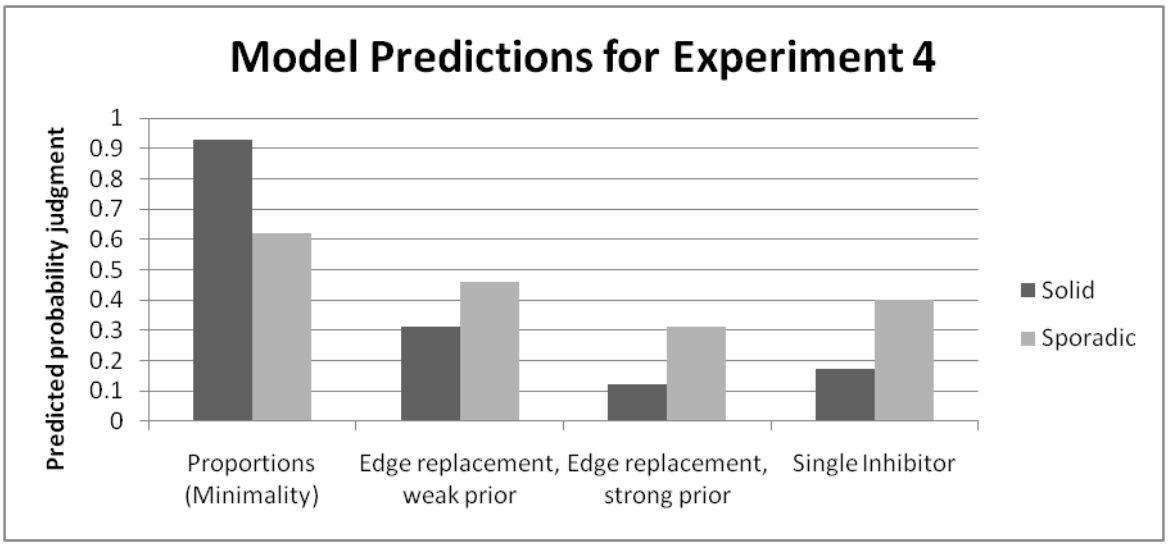

Figure 32: Model predictions for Experiment 4.

Another approach might be to assign more weight to some trials. For instance, a recency component might be justifiable in this scenario, in which recent trials are more heavily weighted when calculating the proportion to be used. This won't help minimality fit the data either: The only way of predicting a low judgment in the solid condition, without predicting a low judgment in the sporadic condition as well, is to assign a high weight to a trial that is a failure in the solid condition, but a success in the sporadic condition. However, no such trial exists: The only failure in the solid condition, is a

failure in the sporadic condition as well. This argument is not restricted to monotonically decreasing recency function: No function, however tailor-made to these results, could come closer to fitting the data, than to predict the same judgment between conditions. Because removing successes actually increased subjects' probability judgments, probabilistic proportional functional forms cannot predict the observed effect.

Similar arguments apply to an alternative approach in which we explicitly model flipping the switch as a relevant intervention on the causal system (Figure 31b). In this case, the minimal model assumes that the state of the switch is part of the functional form of the causal relation. However, minimality has no prior commitments, and data from only one trial, to determine what form the relation should take. Crucially, the trial on which minimality would base its functional form, is identical between conditions. If the switch is relevant, then only the data from after the switch is flipped, is relevant, and the minimal again cannot predict a difference between conditions. Overall, the simplest functional forms within minimal models make predictions that are opposite to the experimental findings. In trying to fit the data post-hoc, the best that more complex minimal models can do, is predict no difference between conditions. This is also inconsistent with the data from 4-year-olds, which showed a significant difference between conditions.

Edge Replacement

Edge replacement makes different predictions. An example of a causal structure, generated by edge replacement, that might fit the data, is shown in Figure 33. Edge

replacement makes a simple qualitative prediction in this case. This section will begin by arguing for this qualitative prediction, before verifying the argument through sampling. Assume that volatility is different in different systems. This is easy to justify: A rock gives much less variable responses than a person, for instance when asked “what time is it” or poked with a stick. We can model this by sampling the volatility of each system randomly. Some of these systems will have low volatility, and some high volatility. Low volatility systems will be more likely to generate the sequence from the solid condition. This same low volatility will make failure especially likely given a failure on the previous trial. Conversely, high volatility systems will be more likely to generate the sequence from the sporadic condition. In these same systems, volatility makes success reasonably likely after a failure, because whatever inhibitor caused the failure, might change back. Applying this generative argument to children’s inferences, children in the solid condition inferred that the system had low volatility, and used this low volatility to predict a failure following a failure.

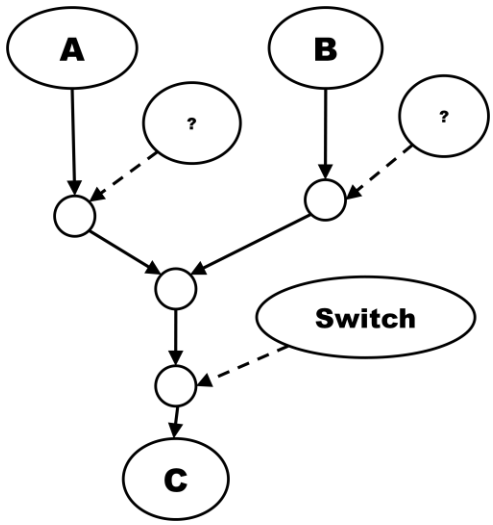

Figure 33: A model generated by edge replacement, that fits the data from Experiment 4.

These predictions were also verified using a sampler. Because inferences where about specific hidden inhibitors, simplified edge replacement could not be used – the full model was required. This necessitated some decisions about parameters. Whenever possible, the same parameters used in other experiments were used. Whenever this was not possible, parameters were sampled from an uninformative prior. Volatility was sampled from $Exponential(1)$, while activation proportions were sampled from Beta(1,1), that is $\alpha = \beta = 1$. All repalcements were assumed to be inhibitory, because the lights never illuminated spontaneously. The sampler set $\gamma = 0.5$, because this fit the data in previous experiments (recall fitting Walsh and Sloman, page 92). The only remaining parameter was λ, the replacement rate of the generative process. To determine the appropriate setting forλ, recall that that *h* = 0.5 fit the data best in previous experiments (again, recall fitting Walsh and Sloman). Because all the parameters together determine h, some unique setting of λ will produce a value of h=0.5, given the other parameters. A series of computational experiments revealed that the optimal setting for λ was 1.6.

For each sample, the program generated a causal structure using these parameters. It then created a single branch point on the graph, because there were exactly two tokens of the cause lights. The branch point was equally likely to occur at any point on the graph, and so was uniformly sampled. The program also sampled the hidden states according to the temporal extension of edge replacement, according to

the volatility parameter for each graph. The sampler generated only inhibitory replacements, because the lights never illuminated spontaneously. The effects of the intervention were modeled as follows: randomly choose an inhibitor from before the branch point and change its state at the 10^{th} time step. That is, if the inhibitor is on, then turn it off, and if it is off, then turn it on. (Graphs that did not have such an inhibitor were rejected, because the intervention could not have occurred.) The state of the inhibitor at the 11^{th} time step was then re-sampled given the new 10^{th} time step, according to the volatility parameter of the graph, and the probability of activation of the inhibitor. The program sampled sequences until it had 10000 samples that matched the sequences from the solid and sporadic conditions, respectively. The predicted probability judgment for the efficacy of each cause light was the proportion of times that that light was active on the 11^{th} trial, among all generated sequences that fit the data.

Model predictions are shown in Figure 32. Edge replacement predicts that children should be more likely to predict efficacy in the sporadic than in the solid condition. This is the converse of minimality's proportional prediction, and fits the data from Experiment 4. The version of edge replacement described above used a weak, minimally informative prior distribution on the volatility values (Exponential(1)), meaning that edge replacement makes this prediction even when it assumes that children are relatively agnostic about the stability of hidden inhibitors. It is possible (even likely) that children have learned that most electronic toys, such as light switches,

are relatively stable in time unless intervened on. For this reason, Figure 32 also shows results for sampler with a stronger prior (Exponential (0.25)) on the distribution of volatility values. These predictions are qualitatively the same: higher probability judgments in the sporadic than in the solid condition. Edge replacement makes the same correct qualitative predictions given these two reasonable parameter settings.

Edge replacement is not the simplest model that can fit these data. One could also argue that children represent a single inhibitor, which evolves in time according to a model similar to edge replacement's temporal component. Figure 32 also shows the results of such a simple model, which generated a sequence of successes and failures, with volatility sampled from Exponential (1). The model treated both cause lights as a single cause, and did not incorporate the intervention. As Figure 32 shows, the results are consistent both with the experimental data and with edge replacement . Because the experiment was designed as a test of edge replacement's predictions, this section presented results from the full model first.

Overall, the most straightforward instances of each class of model, make different predictions: edge replacement makes predictions that qualitatively fit the data, while minimality makes converse predictions that do not fit the data. Extreme versions of each model could be constructed to make different predictions. For instance, we could run edge replacement with a strong prior that prefers volatility in causal systems. Recall that as volatility becomes extremely high, edge replacement's predictions degenerate into proportional predictions. There is no justification for using such a prior:

Causal systems such as the lights are usually relatively stable in the real world. At best, children might not have learned about this stability, but they could not have learned the opposite.

Along the same lines, it is possible that minimality could try to devise a new functional form, specifically to fit the data from this experiment. For instance, we could imagine a time-dependent functional form that did not use a recency function, but instead used a hidden state that changed over time. It is difficult to see how such an approach would differentiate itself sufficiently from the temporal component of edge replacement, to present an alternative explanation. This would amount to fitting the data, by admitting that edge replacement's temporal component was correct. Furthermore, there are aspects of edge replacement's fit that involve other non-minimal commitments, such as branching: 4-year-olds successfully generalized the inefficacy of the right light, to the left light, without having seen the left light's efficacy after the intervention. As we saw in the experiment on stream location, minimality has difficulty explaining this finding, while edge replacement predicts it naturally. Overall, it is not clear how minimality could fit the present data, without leaving minimality behind.

Discussion: Experiment 4

Experimental results from Experiment 4 showed that adding failures to a sequence of trials, one can lower 4-year-olds's judgment of the probability of subsequent success. Four-year-olds were more likely to predict that a causal relation would continue to fail

following an intervention that induced a failure, if the relation had always succeeded (the solid condition) than if it had sometimes failed (the sporadic condition). Three-year-olds failed to show a difference between conditions, or to respond significantly differently from chance on any of the test questions. Modeling results showed that edge replacement could accommodate and predict the 4-year-old findings, proportional models (identified with minimality) could not even fit them post-hoc.

It is unexpected, but interesting, that the data show a developmental difference in this experiment. There are two possible explanations for 3-year-olds' failure. One is that they failed the task because it was too complex. It is possible, for instance, that 3-year-olds lost track of the pattern of successes and failures, and were unable to use it in their reasoning. Further work can investigate this possibility, for instance by using memory aids such as visual reminders of the efficacy of the lights. Another possibility is that 3-year-olds reason about these causal relations in a different way than 4-year-olds. Because 3-year-olds are not responding significantly differently from chance in this experiment, it is difficult to make a strong case from these data about what alternative model they could be using. The simplest, and most likely explanation is that 3-year-olds are simply confused. The conclusion will return to this point, because data from Experiments 5 and 6 is relevant as well.

The data from Experiment 4 provide further support for edge replacement, as an alternative to minimality. Experiments 1-3 showed that edge replacement could be productive, by correctly predicting stream location effects in preschoolers. Experiment 4

shows than another of edge replacement's commitments is productive as well: 4-year-olds seem to posit hidden inhibitors, not just random functional forms, when they encounter a variable causal relation. Further, they seem to reason about the inhibitors in the way that edge replacement predicts. Chapter 5 will discuss further experiments that test edge replacement's predictions about how preschoolers reason about variability, this time with respect to the physical complexity of a system.

CHAPTER 5: VARIABILITY AND COMPLEXITY

The last section of Chapter 2 discussed some of the implications of edge replacement for the theoretical understanding of causal mechanisms. One of these implications was that variability in causal systems was due to inhibitors that are *real:* They have a location and persistence in space and time. Chapter 4 focused on testing the implications of this persistence in time for probability judgments; Chapter 5 will focus instead on the spatial reality of the inhibitors.

The experiments of Chapter 5 present children with causal relations that show either high or low variability. Modeling and experimental results from Chapter 4 suggest that children expect causal systems such as the lights to be composed from hidden causes that each have low or moderate variability. While a system that shows overall moderate or low variability could be constructed using one or two of such hidden causes, a system that shows higher variability would need more such hidden causes, necessarily embedded in a more complex causal structure. This gives rise to the main commitment tested in Chapter 5: Variability implies complexity. Minimality makes no such prediction: using unrestricted functional forms, variability can arise just as easily from simple causal structures as from complex ones.

Because the correspondence between functional and physical space is not yet a completely formalized aspect of edge replacement, this chapter will use less formal detail, and more theoretical arguments. The argument will not rely on a detailed correspondence between functional and physical space, only on the (hopefully

uncontroversial) stance that more often than not, a more complex causal structure corresponds to a more complex physical structure. Experiments 5 and 6, presented in this chapter, will investigate whether children expect additional physical complexity, in cases where edge replacement says they should posit a more complex causal structure.

In Experiment 5, children encountered two toys. The toys were identical, except that one toy instantiated a variable causal relation, while the other toy instantiated a more stable causal relation. Children then saw two pictures of the insides of the toys: one showing a complex inside, the other showing a simple inside. Edge replacement predicts that children should naturally map the complex inside to the variable relation, while minimality predicts that children should have no firm expectations either way.

Experiment 5

Methods

Participants

Participants included 51 preschoolers: 28 3-year-olds (mean age: 41.61 mo, 13 girls), and 23 4-year-olds (mean age: 51.87 mo, 12 girls). Four more children (all three-year-olds) were tested, but were excluded due to experimenter error or equipment failure. Additionally, two children (both 3-year-olds) were excluded because they failed controls. About half of the children were recruited from birth records and tested in a laboratory setting, and about half were recruited at the Providence Children's museum and tested in a quiet room there. The racial and ethnic breakdown was as follows: 47

children were Caucasian, one was Asian, one was African-American, and one was of mixed or other race. No information about socio-economic status was collected.

Materials

Experiment 5 used a modified versions of the smaller lights (the 'cause' lights) from Experiments 1-4. Recall that in previous experiments, the lights were modified such that they illuminated only when actively depressed. In their original, commercially available state, these lights are designed by default to toggle from 'on' to 'off' – if depressed, they illuminate, and continue to illuminate until depressed again. Lights used in Experiments 5 and 6 were left in this original activation state. Two of these lights were modified such that the bulb was replaced with a flashing LED. Each light could be set into a variety of states, but only two were used in this experiment: In one state, the LED illuminated a solid blue color. In the second state, the LED cycled through a sequence of seven colors, approximately once per second. We will refer to these as the 'solid' and 'variable' lights, respectively. The colors, either solid or variable, visibly illuminated the entire top surface of the light. The two lights were otherwise identical: neither were painted, but were left their original white color. A third light was left completely unmodified – it illuminated white using a standard bulb.

Materials also included two fictitious pictures of the 'insides' of the lights. These were made by taking a picture of one of the larger ('effect') lights in a dismantled state, with some of the inside wiring visible, then adding fictitious components that appeared to be part of the workings of the light. The pictures are shown in Figure 34. The 'simple'

picture showed only a light, and a battery, connected by two wires. The 'complex' picture was made by taking a picture of the same scene, but after adding three devices chosen to look vaguely electronic to a child: a voltmeter, a label maker, and a magnetic switch. In reality, the inside wiring of the lights are more complex than the simple picture, and complex in a different way than is shown in the complex picture. No children commented on or apparently resisted the idea that these were pictures of the true insides of the lights.

Finally, a set of simpler stimuli were used in the training phase. These included two transparent boxes, each 5x3.5x3.5 inches, one containing a black triangular wooden block, and another containing a green cubic wooden block. Corresponding to these boxes were two pictures of the insides of the boxes – one a picture of a black triangle, one a picture of a green square.

Procedure

Training Phase

Children were first familiarized with the format of the procedure, using a training phase. The experimenter brought out the two transparent boxes, and arranged them vertically on the table, with one in front of the other. He also put the two pictures horizontally on either side. The arrangement is shown in Figure 34. The arrangement was designed to give no clues as to which picture went with which toy. The experimenter randomly chose whether to put the box containing black triangle in front or back, and also on

which side to put each picture. The experimenter said “We have two different kinds of toys, with two different kinds of insides. And we have some pictures that go with the insides of the toys. Let’s match them up.” He removed the back toy from the table, and pointed to the front toy, with both hands simultaneously. “One of the pictures goes with the insides of this toy. Can you point to the picture that goes with the insides of this toy?” Children who did not point correctly were given corrective feedback . This happened twice – one child succeeded on the second try, and one child was excluded for failing to answer correctly after three tries. Both children were 3-year-olds. At the end of the training phase, the experimenter cleared the table of all the toys.

Test Phase

The experimenter then showed the children the unmodified light, demonstrating to children that when he pressed it, it lit up. Children were allowed to play with the light. The experimenter put the unmodified light away, and brought out the two modified lights, containing LEDs. He said: “We have two different kinds of lights, that have two different insides. Let’s see what they do.” The experimenter depressed each light in turn; the solid light illuminated a solid blue color, the variable light illuminated in a changing sequence. Children were not allowed to touch these lights until the procedure was completed. The experimenter also brought out the two pictures, saying “We have two pictures of the insides of the lights.” The experimenter randomly chose whether to put the solid light in front or behind. The lights and pictures were arranged in the same schema as the training phase, with one difference: The location of the pictures was

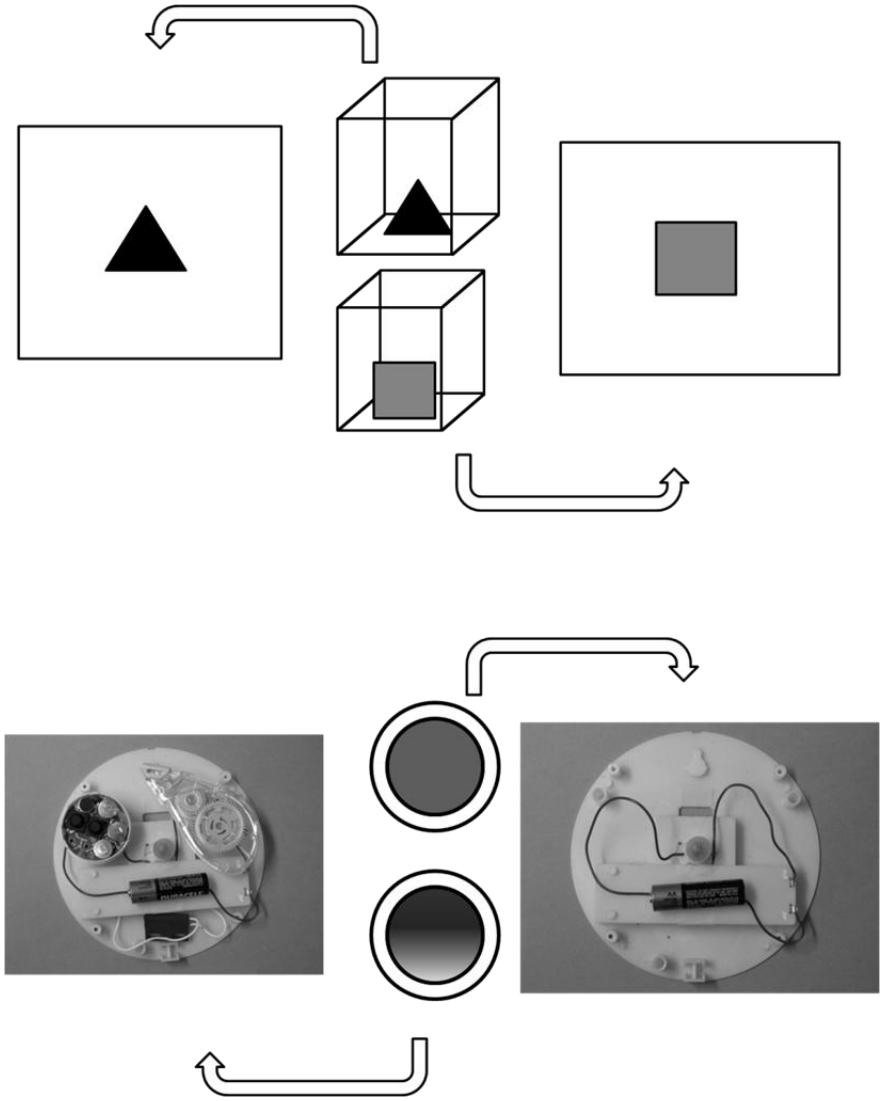

Figure 34: The experimental setup and pictures used in Experiment 5. The arrows (not shown to children) indicate the correct mapping, and the gradient fill indicates the variable light.

determined so that the correct mapping from the training phase was always different from the mapping that would match the variable light to the complex insides (See Figure 34). For instance, if the black triangle was in front in the training phase, and the picture of the black triangle was to the right, (a clockwise mapping) then if the variable light was in front in the test phase, the complex picture was always to the left (a counterclockwise mapping). The experimenter said, pointing to the simple picture: “In this picture, we

have a light and a battery on the inside," then pointing to the complex picture, he said "In this picture, we have a light and a battery and a blicket and a stennet and a gazzer on the inside."

Children were then asked to map the lights to the pictures. The experimenter put the back light away, and said, pointing to the front light with both hands: "One of these pictures goes with the insides of this light. Can you point to the picture that goes with the insides of this light?" Children's responses to this question were recorded and analyzed. Children were also asked to map the back light to a picture.

Results

Most children (51 out of 52) were consistent, never mapping both pictures to the same light. For instance, if they mapped the front light to the right picture, they always mapped the back light to the left picture. The one child (a 3-year-old) who was not consistent, was excluded from the analysis. Because of this, included children fell into one of two categories: Those that mapped the variable light to the complex inside and the solid light to the simple inside (those that 'matched') and those that mapped the solid light to the complex inside, and mapped the variable light to the simple inside (those that did not match.) Overall, 69% (35 out of 51) children matched. This proportion is significantly greater than the proportion (50%) that would be expected if children guessed, and were then consistent, Binomial Test, $p < 0.01$.

Because Experiment 4 showed a developmental difference between 3- and 4-year-olds, each age group was also analyzed separately. Within each age group, no

significant effects were detected for gender, or counterbalanced factors such as the location of the front light or the direction (clockwise or counterclockwise) of the correct mapping. Results are shown in Figure 35. Four-year-olds were significantly different from chance, matching 79% of the time, (18 out of 23 times) Binomial test, $p = 0.011$. Three-year-olds were not significantly different from chance, matching only 61% of the time (17 times out of 28) , Binomial test, p=0.345. No significant difference was found between age groups, Fisher's exact test, p=0.23. Other methods were used to test for a developmental difference. One was to run a Spearman correlation relating matching behavior to age in months. This was not significant, $p = 0.117$. Another method was to use a median split on age at 45 months. In the younger group, 61 % of children (16 out of 26) matched; this was not significantly different from chance, Binomial test, $p = 0.33$. In the older group, 76 % of children (19 out of 25) matched; this was significantly different from chance, Binomial test, $p = 0.014$. The two groups in the median split were not significantly different from each other, Fisher's exact test, $p = 0.37$.

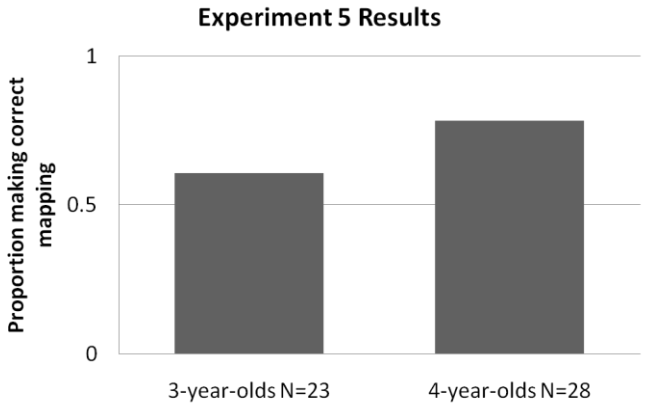

Figure 35: Data from Experiment 5.

Discussion: Experiment 5

Results show that 4-year-olds were able to correctly match a picture of a complex inside, with a variable causal relation. Three-year-olds, on the other hand, were not significantly different from chance. However, there were no significant differences between the age groups, and the entire group of children as a whole showed a significant tendency to match the complex inside to the variable complex relation.

These data are consistent with two interpretations: In one interpretation, 4-year-olds show an ability that 3-year-olds lack. Another interpretation is that both 3- and 4-year-olds share this ability, but 3-year-olds' ability is weaker and more difficult to detect. For instance, 3-year-olds could know that variability implies complexity, but have difficulty with another part of the task. This latter explanation is supported by the fact that some aspects of the experiment are known to be difficult for three-year-olds. For instance, this experiment asked three-year-olds to reason about the insides of objects in a causal context, something they have been previously shown to have difficulty with (Sobel et al. 2007).

A remaining question is exactly what ability 4-year-olds have demonstrated in this experiment. One simple explanation for these data is that 4-year-olds were able to make a heuristic mapping of complexity to complexity. Experiment 6 will test for this by equating the visual complexity of the two lights.

Experiment 6

Experiment 5 showed that preschoolers would match a causally variable light, instead of a causally solid light, to a complex inside. The richest interpretation of these data is that preschoolers understand that causal variability implies internal complexity. An alternative explanation is that children matched the visual complexity of the causally variable light, to the visual complexity of the picture of the complex inside. If this alternative explanation is correct, then children should show chance performance when the two lights show the same level of visual complexity, but only one light shows causal variability. The richer interpretation predicts instead that children should ignore superficial visual complexity, and match the causally variable light to the complex inside in this situation. Edge replacement is committed to this latter prediction, because the hidden inhibitors that instantiate internal complexity should give rise to causal variability, not just superficial visual complexity.

Experiment 6 tested for these predictions by replicating Experiment 5, with one modification: The casing of the solid light was colored using the same variety of colors, with which the variable light flashed. This was intended to equate as much as possible the visual complexity of the two lights. Because the experiment is still underway, a relatively small amount of data has been collected for Experiment 6.

Methods

Participants

Participants included 18 preschoolers, (range: 36-57 months, mean age: 46.17 mo, 3 girls). All children were Caucasian. About half of the children were recruited from birth records and tested in a laboratory setting, and about half were recruited at the Providence Children's museum and tested in a quiet room there.

Materials

The 'solid' light was modified for this experiment: The casing was wrapped in a multicolored band that had been painted in the same seven colors with which the variable light flashed. Materials were otherwise identical to Experiment 5: two lights, one solid and one variable, two pictures, one complex and one simple, and a set of transparent boxes with pictures of their insides.

Procedure

The procedure of Experiment 6 was exactly the same as for Experiment 5, but used materials modified as described above. Children were asked to map the transparent boxes with their insides, then to map the two lights to the pictures of their insides. Children's mappings (whether or not they matched) were recorded.

Results and Discussion

All 18 children passed the training phase (they correctly mapped the transparent boxes to their respective insides) and were consistent in the test phase (they never mapped both lights to the same picture). Because there was insufficient data, age groups were not analyzed separately or compared. The proportion of children who correctly matched the complex light to the causally variable light was 66%, (12 out of 18 children). This was similar to the proportion (69%) that matched in Experiment 5. Because children were above chance in Experiment 5, and the hypothesis was that Experiment 6 would replicate those results, a one-tailed binomial test was performed on these data. This test was not statistically significant, Binomial test, $p = 0.11$. Given that the sample size is still small, the proportionals are similar, and the results of the Binomial test are close to significance, it is reasonable to expect that these data will reach statistical significance, and replicate the results of Experiment 5. More data is needed to be sure.

If this experiment replicates the results of Experiment 5, it will suggest that children are not merely mapping visual complexity to visual complexity. Rather, they are mapping the causal variability of the variable light, to the picture of internal physical complexity, as edge replacement predicts.

General Discussion: Experiments 5 and 6

Experiments 5 and 6 tested the ability of 3- and 4-year-olds to recognize that variability implies complexity, as edge replacement predicts. Experiment 5 showed that 3- and 4-

year-olds would match a more variable causal relation to a more complex inside, when the variable causal relation was more visually complex than its alternative. Experiment 6 is intended to replicate these results when both relations are equally visually complex, but only one is causally complex. Pilot data from Experiment 6 are consistent with Experiment 5.

Experiment 5 found that the responses of 3-year-olds were not, taken by themselves, significantly different from chance performance. However, no significant developmental differences were found, either between 3- and 4-year-olds, or between older and younger children in a median split at 45 months. There is insufficient evidence to show a developmental difference. Grouping all children of all ages together, the data show a significant difference from chance. Taken together, these data do not support the idea that children develop the ability to recognize that variability implies complexity. Rather, it is likely that aspects of the experiment (such as reasoning about internal properties, [Sobel et al. 2007]) confused three-year-olds and made their knowledge more difficult to detect. Further work can focus on detecting this knowledge, for instance by making the procedure rely less on reasoning about pictures of hidden internal properties. For instance, the experimenter could shuffle the lights behind a screen, and open them up, to reveal real and present internal properties.

Because of insufficient data, Experiment 6 did not conclusively rule out an alternative explanation for the data of Experiment 5: That children were matching the visual complexity of the variable light to the visual complexity of the complex picture.

There are other considerations that make this alternative explanation unlikely. For instance, 3-year-olds were not significantly above chance in Experiment 5. Because matching visual complexity is easier than reasoning about internal properties, it is difficult for the alternative explanation to account for 3-year-olds' failure here.

Readers will notice that Experiments 5 and 6 do not line up exactly with the experimental paradigm of Experiment 4. In Experiment 4, variability was expressed by using a light that varied in its success or failure trial by trial. By contrast, Experiments 5 and 6 expressed variability using a light that flashed through a series of colors. In retrospect, it would have been better to conduct Experiments 5 and 6 this way. That is, the experimenter could have shown the children that the 'variable' light sometimes failed when pressed, while the 'solid' light always succeeded. Edge replacement makes the same strong predictions here: The variable light should be matched with the complex insides. Further work can focus on replicating the results of Experiments 5 and 6 using this improved paradigm.

The data from Experiments 5 and 6 support edge replacement's prediction: Preschoolers should know that variability implies complexity, even in a system about which they have no knowledge. Edge replacement prescribes that variability implies the existence of hidden inhibitors. Results from Chapter 4 suggest that preschoolers believe that such inhibitors have low to moderate volatility: they tend not to change state over short periods of time. This means that greater variability implies the existence of more hidden inhibitors. Minimality does not make this prediction, because the unrestricted

functional forms of minimality can account for variability without positing hidden inhibitors. This pattern – that edge replacement produces novel, correct, predictions, while minimality does not – is consistent with the outcome of earlier chapters. The conclusion will review this evidence as a whole.

CONCLUSION

Recap

Over a series of models and experiments, this thesis tested the idea that people represent causal mechanisms using a generative edge replacement rule. The introduction framed the question of mechanism as the representation of causal structure between cause and effect. Chapter 1 introduced Causal Graphical Models, a language in which to address this question. Also in Chapter 1 we encountered the principle of minimality, a rigid preference for simplicity designed to prevent inferring intractably complex causal structures. Reviewing existing data, we saw that minimality failed to fit human data in several ways, including nonindependence effects and an apparent preference for complex deterministic structures. Minimality also seemed theoretically incompatible with representations of causal mechanism. Chapter 2 introduced edge replacement, an alternative way of managing the hypothesis space of possible CGMs, that is friendlier to the idea of mechanism. Edge replacement defines a generative process that can generate any causal relation, and tends to create branching streams. Apparently probabilistic relations are captured by hidden causes that can vary, but tend to be stable in time. Chapters 3,4 and 5 tested some predictions of edge replacement in preschool-aged children. The branching character of edge replacement predicts that children should exhibit stream location effects. Children showed these effects in Chapter 3. Chapter 4 showed that children as young as 4 years old (and

possibly younger) seem to be determinists in the sense that edge replacement predicts: They expect that variability is the result of hidden causes, that persist in time. Chapter 5 showed evidence suggesting that preschoolers also infer that variability implies complexity, as edge replacement predicts. None of minimality's alternative predictions, in any of these experiments, were supported by the experimental data. Edge replacement seems to be a more promising model of human causal reasoning, than models based on minimality. By using a formal model to address a theoretical consideration (mechanism) we discovered an alternative way of looking at causal representations, which made novel, correct predictions. This conclusion will discuss the implications of this work, and directions for further work.

Developmental differences

Before proceeding, it is necessary to address a pattern that might be seen in the experimental data from Chapters 3, 4, and 5. Recall that Experiments 1-3 showed robust, strong stream location effects in both 3- and 4-year-olds. Stream location effects rely primarily on what we will call the 'branching' component of edge replacement: new edges are made from replacements on existing edges. Experiments 4-6 were designed to test a different component of edge replacement: variability is due to stable inhibitors. We will call this the 'determinism' component for simplicity. Experiments 4-6 showed stronger results with 4-year-olds than with 3-year-olds. From this pattern of data, one might be tempted to infer the following hypothesis: By late in the preschool years, children's causal reasoning is captured by edge replacement, but the branching

component develops before the determinism component. While this hypothesis is tempting, this section will present multiple reasons to reject it.

The first reason is that such a hypothesis requires an account of what model younger children *are* in accord with, and no strong evidence exists to suggest what this model is. For instance, recall that Experiment 4 did not show that 3-year-olds reasoned in accord with minimal or proportional models – rather, they were at chance. An experiment is needed in which 3-year-olds respond differently from chance, and from the predictions of edge replacement, but in accord with another model. As of yet, no such experiment exists.

Another piece of evidence is that 3-year-olds show an implicit determinism in the Experiments 1-3. Recall that children saw a series of failures, followed by a change to the mechanism, followed by a success. The difference between conditions in Experiments 1-3, which differed only in the location of the change, indicates that even three-year-olds attributed the change in efficacy to the change made to the mechanism. If three-year-olds truly lacked the determinism commitment of edge replacement, they should attribute the change in efficacy to randomness, not to a change in the mechanism. This would lead to different behavior than we observed – namely, three-year-olds should not show a difference between conditions, because they should not recognize the change as relevant. This is especially true in Experiment 3, where the mechanism was contrary to their existing knowledge (removing a battery enabled the relation), but 3-year-olds apparently still did not attribute the change to chance.

Finally, younger children reason in a way that is much like what one would expect, if they were using something computationally like edge replacement, but becoming confused along the way due to processing limitations. For instance, the deterministic inferences on which 3-year-olds apparently succeed in Experiments 1-3 are easier in many ways than the deterministic inferences made in Experiments 4-6. Experiments 1-3 showed a change from solid failure to solid success – this requires keeping track of only two things. By contrast, the sporadic condition of Experiment 4 required children to keep track of a sequence of nine successes and failures. This pattern of data suggests that 3-year-olds are capable of the inferences edge replacement posits, but sometimes have difficulty applying these inferences to complex data.

Further Experimental work

There are further experiments, motivated by edge replacement, that are likely to yield interesting data. The line of experiments that follows most directly from this thesis is to replicate the findings of Experiments 1-5 with adults. Some modifications will be necessary -- it is not reasonable to expect that adults will answer in exactly the same ways as children given exactly the same stimuli. For instance, recall that in Experiment 3, batteries were removed from a toy, an action which then enabled its efficacy, contrary to the way toys actually operate. Adults would be especially likely to suspect a trick in this case, and so might give a different response than children. Similarly, Experiment 5 showed children pictures of the insides of lights that were designed to resonate with a

childlike level of electronics knowledge. Several parents spontaneously reported that they were unsure which inside went with each light, even when their child made the correct mapping (that is, mapping, the variable relation to the complex insides).

The key to replicating these results with adults, would be to ask adults to reason about a causal system about which they are similarly ignorant. For instance, we have seen that Mayrhofer et al (2010) used a cover story about alien telepathy, which was novel enough to elicit phenomena we explained using stream location effects. One approach to replicating Experiments 1-4 would be to borrow this paradigm: Instead of a light, whose relations were enabled when a battery was removed, we could tell adults about an alien, whose ability to concentrate is enabled by taking medication. It would be important to replicate Experiments 1-3 this way, despite the fact that Mayrhofer et al (2010) already show results that suggest stream location effects. One reason for this is that the predictions are more direct: Mayrhofer et al told subjects only about transmission or reception, rather than directly telling them about the location of the inhibitor – edge replacement accounts for the difference they observed in nonindependence effects by assuming that participants inferred the location of the inhibitor from this description. By directly manipulating the location of the inhibitor, we can bypass this additional interpretive step. There is also merit in testing a novel prediction, rather than explaining existing data post hoc.

Similar logic applies to other experiments: We could replicate Experiment 5 by discussing a machine that achieves an outcome that adults know is impossible given

current technology, such as a teleporter or a machine that makes gold out of cheese. Adults should infer more complex causal structure for variable machines than for machines that show a more stable effect. Because it contains a structural prior even under uncertainty, edge replacement makes its strongest and clearest predictions about reasoning involving systems about which the participant knows very little. This is one of the reasons that children were chosen for initial tests – there are many systems about which they know very little. Analogous adult experiments will involve recreating this uncertainty.

Further experiments with adults would be able to test and develop aspects of edge replacement that developmental data have not been able to address. For instance, it is easier to elicit clear and reliable probability strength judgments from adults – these could be used to test some of the more subtle quantitative predictions of the model. We can also more readily explore parametric variations on the experiments discussed above. For instance, in a version of Experiment 5, we could continuously vary the amount of complexity present in the pictures, and also the amount of variability present in the causal relations. Edge replacement should be able to predict how this will affect participants' mappings, in more subtle ways than can be detected in a developmental experiment. In a version Experiment 4, we could vary both the amount of variability present in the sequence of successes and failures. The 'solid' and 'sporadic' sequences of Experiment 4 are only two ends of a continuum; using longer sequences, we could

test intermediate amounts of variability, and see if people infer intermediate amounts of volatility. Edge replacement can make clear quantitative predictions here.

Working with adults also gives us an opportunity to test predictions that are difficult to test in children, even qualitatively. One example of such a prediction has to do with the number of hidden causes that are present in a causal system. Edge replacement assumes that hidden causes (that is, replacements) are equally likely to exist at any given point in a causal stream. We can use this assumption to generate a prediction using the following line of reasoning: The earliest of ten random numbers is likely to be lower, on average, than the earliest of three random numbers. This implies that the first of ten inhibitors will be earlier in the causal system than the first of three inhibitors. Earlier hidden causes create stronger nonindependence effects when described as prone to failure. Together, these aspects of the model imply that if we vary the number of described inhibitors in a causal system, we may be able to change the degree of nonindependence observed when the earliest cause is described as prone to failure. Edge replacement should be able to quantitatively predict these experimental effects.

Further developmental data could also address aspects of edge replacement, or provide further empirical support for the model. Some of these possible experiments were discussed in Chapter 4. For instance, it should be possible to replicate Experiments 1-3 using a common cause scenario, instead of the common effect scenario that was used there. An open question from Experiments 4 and 5 was whether there is a version

of the task on which 3-year-olds might show above-chance performance, rather than the chance level of performance observed there. We might be able to achieve this by making the task easier, for instance by shortening the sequences about which children need to reason, or redesigning tasks such that children do not need to reason about the hidden internal properties of objects.

Another possible extension would be to further investigate the results of Experiment 4. Recall that we assumed that children had some prior distribution on the volatility parameter. We may be able to experimentally manipulate this prior expectation of volatility. For instance, in a training phase, we might show children that in general, the lights tend to be unstable, spontaneously changing state. According to edge replacement, this should change their prior expectation of volatility, and make them more likely to predict success in the solid condition.

Experiment 5 could also be extended. In the data presented in this thesis, we chose to instantiate variability using a light that flashed autonomously in a series of colors, once it has been pressed. As discussed at the end of Chapter 5, it would have been more in line with previous experiments to have the variable light sometimes fail when depressed, as in the solid condition of Experiment 4. Another approach might have been to press the light multiple times, and have it illuminate a different color each time, without flashing. In this alternate experiment, the solid light would also be pressed multiple times, but illuminate the same color each time. If edge replacement is correct, then this alternative experiment should replicate the results of Experiment 5.

Other developmental experiments might replicate some of the present findings with younger children. Because even three-year-olds showed strong stream location effects, it is reasonable to suspect that we might find stream location effects in younger children or even infants. It is easy to imagine a version of the stream location experiment, for instance, that uses a habituation paradigm: If both relations in a common effect fail, and an intervention close to one of the causes restores one of the relations, infants should be surprised if this change also restores the other relation. They should not be surprised at this outcome, however, if the change was close to the common effect.

The above experiments are all extensions of the model and experimental paradigms presented in Chapters 1-4. Far more experimental ideas are likely to arise from the further development of the model, a question we turn to now.

Further Modeling work

Edge replacement is not intended to function by itself as a complete model of causal reasoning. Rather, it is intended as one amendment (presumably among many) to help a larger research program better model and understand human representations of cause and effect. As we will see in this section, much of the further work suggested by edge replacement involves making contact between these models.

The relation to space, objects, and force

As mentioned in Chapter 3, edge replacement presumes a mapping between the functional space described by the graphs it generates, and the physical space humans encounter in the world. In the experiments described in Chapters 3-5, we relied only on the most uncontroversial aspects of that mapping. For instance, Chapter 3 assumes that there is a preservation of order: Functionally prior events must be physically prior and vice versa. Further work can explore further aspects of this mapping.

One aspect that is particularly open for further work is the relation to objects. Edge replacement deals not with objects but with events (such as actions on objects) situated in functional space. Many theories of causation (e.g. White, 1989) posit that in causal representations, it is these objects that have causal power. Further work can focus on the interface between such theories and edge replacement. One possible avenue is to see an object as a *cluster* of actions and events. That is, the mapping between physical and functional space, would be strongly biased to assign events related to the same object, to proximate locations in functional space.

Another aspect for further work is the relation to notions of force. Multiple theorists (e.g. Shultz, 1982) and models (e.g. Wolff, 2007) posit that representations of causation operate primary by imagining the transmission of force or energy within the physical environment. Such models are particularly adept at modeling the results of lower-level physical interactions (such as boats being affected by wind, or billiard balls on a table); they are less directly useful for understanding how people represent more

abstract or invisible relations, such as how smoking causes cancer, or how one cellphone causes another to ring. Edge replacement is designed to represent these latter, higher-level relations, but is deliberately constructed to maximize the opportunities for making contact with lower level models. Further work can explore and develop this interface. For instance, Chapter 2 argued that edge replacement lends itself naturally to thinking of force flowing through functional (and physical) space in a continuous manner, being stopped and started along the way by inhibitors and generative causes. This is much more in line with force-dynamical models than the alternatives presented by minimality.

Edge replacement as a learning mechanism

Most of this thesis has discussed edge replacement as a mechanism for generating representations of causal structure within the time scale of a single experimental session. Edge replacement might also be able to account for how causal representations change over developmental time. Consider an apparent computational problem in the development of causal reasoning: As the number of candidate causes increases, examining all possible interactions of causes creates an intractably large combinatorial explosion of possibilities. For instance, imagine an infant who has noticed that when she cries, her mother feeds her. This relation does not always hold – sometimes crying does not immediately yield anything. The infant might choose to represent the relation as probabilistic. If she does so, then there is little or no motivation to explore the possibility that other variables (for instance, if mother is on the phone) are affecting the outcome of the intervention. Assuming that she is interested in a deeper understanding, any

number of things could be relevant: Whether the light in her room is on, whether it is raining outside, what clothing the infant is wearing, whether it is Thursday, etc.

Edge replacement suggest a reasonable search strategy in such situations: Begin with the causal relation of interest. Then posit a hidden inhibitor, and a hidden generative cause, to explain the variability observed. Then search your environment for variables that match these patterns of activity. The match may not be perfect, but finding a candidate cause that explains part of the variability, allows you to grow your model. Once this happens, new questions open up: you can look for further variables that explain the remaining variability. Notice in this situation that because of the restricted functional forms of edge replacement, variability drives a search for deeper understanding. Minimality does not provide such an account of what drives a search for a deeper understanding. Edge replacement predicts that we should devote the most attention to causal relations that are 1) of interest and 2) contain interesting variability.

It will be worthwhile to work through this idea using an example. Consider a causal relation between A and B, where B is some outcome of interest. In reality, there are three other variables, C, D, and E, that are relevant, hidden among many that are not relevant. See the rightmost graph in Figure 36 for a visual illustration. C is a preventer of the AB relation, D is an alternative generative cause, and E is an event that prevents C from preventing the AB relation. To ground the example, imagine a car that will not start when the electrical system is moist. Imagine also that the car has a leaky hood (an example from the author's experience). A jump start (D) will always start (B)

the car, but turning the key (A) usually does, unless it is raining (C). But if the car is indoors or has been parked under a tree (D), rain will not disable the relation.

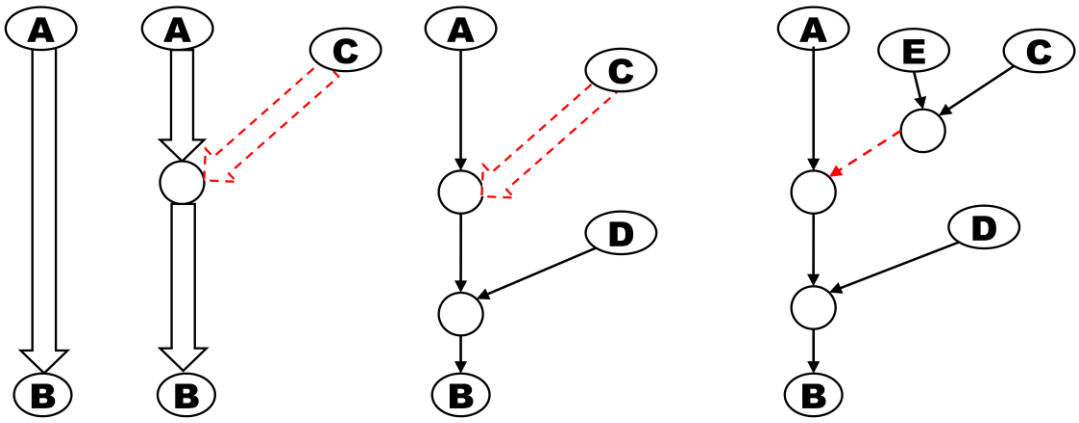

Figure 36: A process through which edge replacement might be used to learn about four variables. Open arrows indicate relations that contain unexplained variability.

Figure 36 shows how edge replacement would account for the way in which this complex structure was learned. The learner begins by searching out generative and inhibitory causes. (C and D). Then, when variability still remains, the learner searches for additional hidden causes in the relations that still show the most variability (namely the edge between C and the bridge node that connects it to the main path). Consequently, edge replacement predicts where learners should focus their attention at different stages of learning, and what order the sequence of attention should follow (focus on C and D should precede focus on E). It should be straightforward to design a behavioral experiment to test these predictions.

Learning edge replacement from simpler rules

Besides its application as a learning mechanism, it is possible that edge replacement might itself be learned. For instance, Goodman, Ullman, and Tenenbaum (2010) have

suggested and demonstrated that theories of causality might be deducible from simpler, underlying rules. Kemp and Tenenbaum (2008) have also shown that by using a 'meta-grammar' – a set of rules out of which grammars can be created – we can explain some of the steps that children make when learning about the categorical structure of the world. Kemp and Tenenbaum were working with node replacement, rather than edge replacement, grammars, but it is possible that a similar approach might work with edge replacement: There might be a set of underlying rules out of which edge replacement could be learned. For instance, children might at first entertain a broader range of functional forms, or a different replacement rule.

One way of approaching this problem would be to consider two meta-hypotheses: Minimality, and edge replacement. We could assume that both of these hypotheses are present in children's very early cognitive development. Modeling work might be able to establish that there is sufficient evidence in children's environment to reject minimality in favor of edge replacement. That is, we could establish that the probability of edge replacement given the data is greater than the probability of minimality given the data. This work might be an important first step, but to stop there would be disingenuous. It is much more important to specify where the two hypotheses come from.

The most coherent way to approach the question of whether edge replacement might be learned, is to look at it within the context of other models, for instance models of categorization. Kemp and Tenenbaum's work, in particular, suggests that children

may be equipped with cognitive mechanisms that instantiate a set of relatively simple, highly generative rules. Over time, the structures generated by these rules settle into theories and grammars in a hierarchical fashion. Human beings learn not just facts, and not just theories, but also new ways of learning, that transfer within and between domains. We can think of this process metaphorically as a representational toolbox – when encountering a problem, children try a variety of tools from the toolbox, (including tools that make other tools) and continue using whatever works. Kemp and Tenenbaum show that their meta-grammatical algorithm will tend to first try simple structures (like relational trees and chains), and if these don't work, move on to more complex structures, such as helices (like DNA) or cylinders (like the periodic table).

One way of incorporating edge replacement into this more hierarchical framework, would be to introduce edge replacement as part of an expanded initial toolbox. It may be that children start with the ability to use many types of grammars, including node replacement, edge replacement, and possibly others (even minimal approaches.) This thesis suggests that edge replacement works well for representing causation, but children may not know this at first – they may try node replacement or even minimal approaches, before discovering that edge replacement fits the branching character of causation. Conversely, children may try edge replacement in attempting to understand categorical structure, and reject it in favor of a node replacement approach. Further work may investigate this formal comparison explicitly. Experimental work could focus on trying to detect early instances in which children or infants might try to apply

the wrong sort of meta-grammar to a certain domain (edge replacement in categorization, for instance) or conversely to find instances in which children outperform adults because the solution to a problem involves a meta-grammar that is not usually effective in that domain.

Understanding the exact nature of children's representational toolbox, along with the rules for its elaboration is well outside the scope of this dissertation. Edge replacement may be a piece of the puzzle. The main argument of this section is that we are better served by looking at causation not in isolation, but as part of the broader learning that is going on in early cognitive development. This 'unite and conquer' strategy may be most appropriate for understanding very early representations.

The question of implementation

Marr (1982) defines three levels of analysis in cognitive science: the *computational* level deals with the problems that are being solved, the *algorithmic* level deals with the steps that are actually carried out in solving the problem, and the *implementational* level deals with the actual mechanism (often the physical substrate) on which the algorithms are carried out. Edge replacement is intended as a computational level model, and so does not carry theoretical commitments about how it is carried out algorithmically or implemented. However, it is important that a good computational level model at least have a plausible story about how it might be carried out at lower levels. If edge replacement happens, it happens in brains.

One of edge replacement's strengths is that it seems algorithmically realistic. This is one of the reasons that a generative grammar was chosen. A problem for some Bayesian models is that an infinite hypothesis space, while formally coherent, seems algorithmically unrealistic. In a finite brain, there is no room for an infinite hypothesis space. Generative grammars are one of many ways of making such spaces tractable: they provide a principled way of navigating the space of possibilities, while considering only small part of the hypothesis space at a time. For instance, one possibility is that edge replacement captures the mechanism through which specific representations of specific systems are generated: When encountering a causal system, some part of the mind generates causal structures using edge replacement, until something is found that fits. The first representation that fits, is used in reasoning. Over multiple subjects, the effect of such a process would resemble Bayesian inference over the infinite hypothesis space.

More mysterious is the question of how edge replacement might be implemented in a brain. This is an open question, and attempting to address in detail would be highly speculative and premature. The key problem is that edge replacement requires structured, relatively logical representations, rather than just associations. This is a problem not just for edge replacement, but for any theory that posits representations that go beyond association. At present, association is better understood at the implementational level. This does not mean that association is all that brains can do: Several other phenomena, such as language, show that the brain manifestly goes

beyond association, at least in some cases, so this requirement is not a reason to reject edge replacement. Instead, edge replacement is one small phenomenon that adds to the importance of understanding how the brain implements representations of richly structured phenomena.

Edge replacement as a computational tool

The question of implementation raises an issue that has not been discussed in this thesis: Whether edge replacement is computationally as tractable as minimality for implementation in digital computers. While this question is not directly related to edge replacement as a psychological model, it is peripherally related to the plausibility of implementing edge replacement in the brain. Edge replacement may seem at first to be prohibitively computationally complex, because the graphs it generates have more nodes and edges than those generated by minimality. However, there are several reasons to suspect that edge replacement may be more tractable in many situations. The first reason is that the restricted functional forms employed by edge replacement are natural for computers to work with – indeed the AND NOT relation is fundamental to computer architecture at low levels. While the graphs that edge replacement works with may be more complex, it may achieve a countervailing decrease in complexity when working with the restricted functional forms. Another reason is that edge replacement tends to produce binary branching tree structures, a natural and efficient data structure for computers to work with. In fact, it is possible to implement edge replacement using a single recursive class definition. Computers and brains are not the

same, but ease of implementation in computers provides support and plausibility to the possibility of implementation in brains.

Is edge replacement rational?

Anderson (1991) describes a rational analysis as a process that can be productively repeated: We begin with mathematical assumptions about what is rational, along with the constraints present in the mind and the environment, then test to see whether people reason in accord with these assumptions. When people fail to reason in accord with these predictions, there are two possibilities: Either people are reasoning suboptimally, or our assumptions about rationality are incorrect. We can see nonindependence and determinism phenomena as an instance in which people failed to reason in accord with the assumptions of a particular theory of rationality, namely minimality. Edge replacement fits human data better; this raises the question of whether edge replacement is intended as a model of people's suboptimal reasoning, or an amended, improved model of what is rational, deduced from human behavior.

Edge replacement can be seen as alternative rational model of causation. This does not mean that minimality is metaphysically incorrect (though data suggest it is psychologically incorrect), only that the assumptions and predictions of edge replacement can be rationally justified. In this sense, edge replacement constitutes a reasonable alternative to minimality. Furthermore, there are at least some situations in which it clearly outperforms minimality, in the sense that it more quickly leads to a

more useful representation of the world. These are, of course, examples in which minimality and edge replacement make the most divergent predictions: Cases of nonindependence, and cases of complex deterministic systems.

As we have seen, minimality initially assumes that collateral events, such as collateral effects of a cause, are independent. Only as statistical evidence accumulates does minimality license a modified structure. Edge replacement, on the other hand, initially expects a moderate correlation between collateral events, due to the branching structure that it requires. From edge replacement's point of view, minimality's independence assumption amounts to an extreme on the prior distribution of branch points: Assume that all collateral effects are generated from very early in the causal stream that connects the cause and effects. This is much like going to Las Vegas, and assuming by default that every game is heavily biased in your favor. For each game, over time, you would gradually learn that the game was approximately fair, or slightly biased against you. While you would eventually converge on the correct probabilities, this process would cost you a great deal of money. Similarly, minimality's persistent assumption that collateral effects are independent may lead to a high cost in terms of predicted independence relations. The primary motivation for assuming minimality is computational convenience, a convenience which leads to models that are consistently wrong.

Edge replacement outperforms minimality, then, in situations where collateral effects tend to be correlated. This raises the difficult issue of what causation is 'really

like' in the world – whether collateral effects really are independent more often than not. This question is difficult, if not impossible, to answer definitively, because it would be almost impossible to conduct a satisfying experiment on the matter. We would have to decide on a population to sample from – do we focus on the kinds of causal relations that people encounter in their everyday lives? Or on all the causation that exists in the universe? Just as most living things are bacteria, most causation happens at the molecular or quantum-mechanical level, creating a strange skew in our potential sample. We must rely instead on argument and experience. Cartwright (1999) provides a series of examples in which causal relations in the world persistently violate simple independence assumptions. For instance, one of Cartwright's examples is a factory produces both useful chemicals and pollution. We might think that the absence of pollution on a given day suggests that the factory is less likely to produce chemicals on that day. This inference is intuitive, and arguably rational, even if we know nothing about the specific process through which the chemicals are produced. Minimality says that the inference is unjustified. We have also seen several examples on nonindependence (Walsh and Sloman's [2008] jogging example, Mayrhofer et al's [2010] mind-reading aliens) in which the initial prediction of nonindependence seems reasonable. Certainly we must accept that at least in some domains, nonindependence occurs more often than independence. Edge replacement is more rational than minimality in those domains, because it begins closer to the truth.

A second respect in which edge replacement may be more rational is that the kinds of representations that it ultimately produces may be more useful. Minimality tends to produce simple, probabilistic structures, that are streamlined for computational convenience, while edge replacement tends to produce more complex, deterministic structures. There are clearly cases in which the latter are preferred. For instance, imagine a robot, programmed only with minimality, that learned to be an auto mechanic. Your car is obviously having a problem, because it only starts 5 days out of 10. You take it to the minimal-robot-mechanic, who diagnoses your problem by saying: "Your car starts with probability 0.5." While strictly correct, this analysis is distinctly unhelpful. You take your car to a human mechanic for a second opinion, and she tells you "You need a new alternator." Edge replacement is more likely to produce representations that lead to this more helpful inference, and so is more rational for these applications.

The primary motivation for minimality is computational: It solves the problem that there are an infinite number of complex representations for any given causal relation. As the evidence from Chapters 3, 4 and 5 shows, the human brain seems to solve this problem in a different way. Edge replacement may not describe with complete accuracy how people do this, but it comes closer than minimality. Rather than imposing the draconian simplicity of minimality, edge replacement provides a way of navigating through the space of more complex (and potentially more useful),

deterministic structures, in a computationally realistic way. In that sense, it is a rational model.

Conclusion

This thesis began by promising to shed light on the question of mechanism: whether and how people represent what happens between cause and effect. One position in this debate is that people do not represent mechanisms when representing causal relations; Causal Graphical Models with minimality were situated in this camp. The alternative position, namely that people represent mechanisms, did not have a similarly specific and formal model allowing it to make predictions. The introduction promised that edge replacement would fill this gap, allowing us to more directly compare the predictions of the two theories.

Now that we have looked at the description, extension, and application of edge replacement, and experimentally tested the predictions of both edge replacement and minimality, it seems that a mechanism-dependent account of causal relations is a clear winner. Moreover, the formal account developed along the way sheds light on the broader question with which this dissertation began: How human beings develop representations of complex causal structure. Edge replacement suggests that when encountering novel causal phenomena, people are not minimalists: instead, they infer a rich, deterministic, causal structure. This may explain better how we go from simple

representations to complex ones: We are driven to more complexity by the nature of our causal representations.

BIBLIOGRAPHY

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*(3), 299-352.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, 41(4), 361-416.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409-429.

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York: Academic Press.

Carroll, C., & Cheng, P., (2010) The induction of hidden causes: Causal mediation and violations of independent causal influence. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 913-918). Austin, TX: Cognitive Science Society.

Cartwright, N. (1999). Causal diversity and the Markov condition. *Synthese, 121*(1), 3-27.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7), 335-344.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*(2), 367-405.

Chomsky, N. (1957) *Syntactic Structures.* Mouton.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines, 8*(1), 39-60.

Glymour, C. N. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge: MIT Press.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25*(4), 565-610.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2009). Learning a theory of causality. In N. Taatgen & H. van Rijn, (Eds.) *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2188-2193). Austin, TX: Cognitive Science Society.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*(1), 3-32.

Gottfried, G. M., & Gelman, S. A. (2005). Developing domain-specific causal-explanatory frameworks: The role of insides and immanence. *Cognitive Development, 20*(1), 137-158.

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik and L. Schulz, (Eds.) *Causal Learning: Psychology, Philosophy, and Computation*, pp 323–345. Cambridge: MIT Press.

Heather Fritzley, V., & Lee, K. (2003). Do young children always say yes to yes–no questions? A metadevelopmental study of the affirmation bias. *Child Development, 74*(5), 1297-1313.

Heckerman, D., Meek, C., & Cooper, G. (2006). A Bayesian approach to causal discovery. in D E. Holmes and L. C. Jain ,(Eds.) *Innovations in Machine Learning: Theory and Applications,* pp 1-28. New York : Springer, 2006.

Hume, D (2000) *An enquiry concerning human understanding: a critical edition.* Oxford University Press. Original work published 1748.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, 7(8), 368-373.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687.

Laplace, P. S. (1951). *A philosophical essay on probabilities*. (F. W. Truscott & F. L. Emory, Trans.). New York: Dover. (Original work published 1814).

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development, 81*(3), 929-944.

Lewis, D. (1973). Causation. *The Journal of Philosophy, 70*(17), 556-567.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*(4), 955–984.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co., Inc. New York, NY, USA.

Mayrhofer, R., Hagmayer, Y., & Waldmann, M., (2010) Agents and causes: A Bayesian error attribution model of causal reasoning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 925-930). Austin, TX: Cognitive Science Society.

McClelland, J. L., & Thompson, R. M. (2007). Using domain-general principles to explain children's causal reasoning abilities. *Developmental Science, 10*(3), 333-356.

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development, 5*(2), 193-207.

Pearl, J. (1988) *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the stimulus? A rational approach. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 663-668). Austin, TX: Cognitive Science Society.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*(3), 264-314.

Reichenbach, H., (1956) *The Direction of Time*. Berkeley and Los Angeles: University of California Press.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64-99). New York: Appleton-Century-Crofts.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521-562.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press: Princeton, NJ.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development, 77*(2), 427-442.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development, 47*(1), 1-51.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives.* Oxford University Press.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*(3), 303-333.

Sobel, D. M., Yoachim, C. M., Gopnik, A., Meltzoff, A. N., & Blumenthal, E. J. (2007). The blicket within: Preschoolers' inferences about insides and causes. *Journal of Cognition and Development, 8*(2), 159-182.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Cambridge, Mass.: MIT Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search.* MIT Press.

Walsh, C. & Sloman, S. A. (2008). Updating beliefs with causal models: Violations of screening off. Gluck, M. A., Anderson, J. R , & Kosslyn, S. M. (Eds.). *Memory and Mind: A Festschrift for Gordon H. Bower.* (pp. 345-357) New Jersey: Lawrence Erlbaum Associates

Watson, J. (1979). Perception of contingency as a determinant of social responsiveness. In Thoman, E.B., (Ed), *Origins of the infant's social responsiveness* (pp 33–64). Lawrence Erlbaum Associates.

Weed, D. L., & Hursting, S. D. (1998). Biologic plausibility in causal inference: Current method and practice. *American Journal of Epidemiology, 147*(5), 415.

White, P. A. (1989). A theory of causal processing. *British Journal of Psychology*, 80(4), 431-454.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology, 136*(1), 82.

Yuille, A., & Lu, H. (2007). The noisy-logical distribution and its application to causal inference. *Advances in Neural Information Processing Systems*, 20