Integrated Inferences

Macartan Humphreys and Alan Jacobs

Draft!: 2020-10-17

Contents

Preface

Quick Guide

This book has four main parts:

- Part I introduces causal models and a Bayesian approach to learning about them and drawing inferences from them.
- Part II applies these tools to strategies that use process tracing, mixed methods, and "model aggregation."
- Part III turns to design decisions, exploring strategies for assessing what kind of data is most useful for addressing different kinds of research questions given knowledge to date about a population or a case.
- Everything up to Part IV assumes that we have access to models we are happy with. In Part IV we turn to the difficult question of model justification and outline a range of strategies on can use to justify causal models.

We have developed an `R` package `CausalQueries` that accompanies this book, hosted on Cran. In addition, a supplementary Guide to Causal Models serves as a guide to the package and provides the code behind many of the mode models used in this book.

Chapter 1

Introduction

We describe the book's general approach, and explain how it differs from current approaches in the social sciences. We preview our argument for the utility of causal models as a framework for choosing research strategies and drawing causal inferences from evidence.

The engineer pressed the button, but the light didn't turn on.

"Maybe the bulb is blown," she thought.

She replaced the bulb, pressed the button and, sure enough, the light turned on.

"What just happened?" asked her philosopher friend.

"The light wouldn't turn on because the bulb was busted, but I replaced the bulb and fixed the problem."

"Such hubris!" remarked her friend. "If I understand you, you are saying that pressing the button *would have* caused a change in the light *if the bulb had not been busted*."

"That's right."

"But hold on a second. That's a causal claim about counterfactual events

in counterfactual conditions that you couldn't have observed. I don't know where to begin. For one thing, you seem to be inferring from the fact that the light did not go on when you pressed the button the first time that pressing the button the first time had no effect at all. What a remarkable conclusion. Did it never occur to you that the light might have about to turn on anyway—and that your pressing that button at just that moment is what *stopped* the light going on?"

"What's more," the philosopher went on, "you seem also to be saying that pressing the button the second time *did* have an effect because you saw the light go on that second time. That's rather incredible. That light could be controlled by a different circuit that was timed to turn it on at just the moment that you pressed the button the second time. Did you think about that possibility?"

"On top of those two unsubstantiated causal claims," the philosopher continued, "you are *also* saying, I think, that the difference between what you *believe* to be a non-cause on the first pressing and a cause on the second pressing is itself due to the bulb. But, of course, countless other things could have changed! Maybe there was a power outage for a few minutes."

"That hardly ever happens."

"Well, maybe the light only comes on the second time the button is pressed."

"It's not that kind of button."

"So you say. But even if that's true, there are still so many other possible factors that could have mattered here—including things that neither of us can even imagine!"

The philosopher paused to ponder her friend's chutzpah.

"Come to think of it," the philosopher went on, "how do you even know the bulb was busted?"

"Because the light worked when I replaced the bulb."

"But that means," the philosopher responded, "that your measurement of the state of the bulb depends on your causal inference about the effects of the button. And we know where that leads. Really, my friend, you are lost."

"So do you want me to put the old bulb back in?"

1.1 The Case for Causal Models

In the conversation between the philosopher and the engineer, the philosopher disputes what seems a simple inference. Some of her arguments suggest a skepticism bordering on paranoia and seem easily dismissed. Others seem closer to hitting a mark: perhaps there was nothing wrong with the bulb and the button was just the kind that has to be pressed twice.

While the philosopher's skepticism guards against false inferences, it is also potentially paralyzing.

The engineer brings background knowledge to bear on a question and deploys causal models of general processes to make inferences about particular cases. The philosopher brings a skeptical lense and asks for justifications that depend as little as possible on imported knowledge.

Social scientists have been shifting between the poles staked out by the philosopher and the engineer for many years. This book is written for would-be engineers. It is a book about how we can mobilize our background knowledge about how the world works to learn more about the world. It is, more specifically, a study in how we can use causal models of the world to design and implement empirical strategies of causal inferences.

There are three closely related motivations for our move to side with the engineers. One is is a concern over the limits of design-based inference. A second is an interest in integrating qualitative knowledge with quantitative approaches, and a view that process tracing is a model-dependent endeavor. A third is an interest in better connecting empirical strategies to theory.

1.1.1 The limits to design-based inference

The engineer in our story tackles the problem of causal inference using models: theories of how the world works, generated from past experiences and applied to the situation at hand. The philosopher maintains a critical position, resisting models and the importation of beliefs not supported by evidence in the case at hand.

The engineer's approach recalls the dominant orientation among social scientists until rather recently. At the turn of the current century, multivariate regression had become a nearly indispensable tool of quantitative social sci-

ence, with a large family of statistical models serving as political scientists' and economists' analytic workhorses for the estimation of causal effects.

Over the last two decades, however, the philosophers have raised a set of compelling concerns about the assumption-laden nature of standard regression analysis, while also clarifying how valid inferences can be made with limited resort to models in certain research situations. The result has been a growth in the use of design-based inference techniques that, in principle, allow for model-free estimation of causal effects (see **?**, **?**, **?**, **?** among others). These include lab, survey, and field experiments and natural-experimental methods exploiting either true or "as-if" randomization by nature. With the turn to experimental and natural-experimental methods has come a broader conceptual shift, with a growing reliance on the "potential outcomes" framework as a model for thinking about causation (see **?**, **?** among others) and a reduced reliance on models of data-generating processes.

The ability to estimate average effects and to calculate p-values and standard errors without resort to models is an extraordinary development. In Fisher's terms, with these tools, randomization processes provide a "reasoned basis for inference," placing empirical claims on a powerful footing.

While acknowledging the strengths of these approaches, we also take seriously two points of concern.

The first concern—raised by many in recent years (e.g., **?**)—is about design-based inference's scope of application. While experimentation and natural experiments represent powerful tools, the range of research situations in which model-free inference is possible is inevitably limited. For a wide range of causal conditions of interest to social scientists and to society, controlled experimentation is impossible, and true or "as-if" randomization is absent. Moreover, limiting our focus to those questions for, or situations in which, exogeneity can be established "by design" would represent a dramatic narrowing of social science's ken. It would be a recipe for, at best, learning more and more about less and less. To be clear, this is not an argument against experimentation or design based inference; yet it is an argument for why social science needs a broader set of tools.

The second concern is more subtle. The great advantage of design-based inference is that it liberates researchers from the need to rely on models to make claims about causal effects. The risk is that, in operating model-free,

researchers end up learning about effect sizes but not about models. But models are what we want to learn about. Our goal as social scientists is to have a useful model for how the world works, not simply a collection of claims about the effects different causes have had in different times and places. It is through models that we derive an understanding of how things might work in contexts and for processes and variables that we have not yet studied. Thus, our interest in models is intrinsic, not instrumental. By taking models, as it were, out of the equation, we dramatically limit the potential for learning about the world.

1.1.2 Qualitative and mixed-method inference

Recent years have seen the elucidation of the inferential logic behind "process tracing" procedures used in qualitative political science and other disciplines. In our read, the logic provided in these accounts depends on a particular form of model-based inference.[1]

While process tracing as a method has been around for more than three decades (e.g., **?**), its logic has been most fully laid out by qualitative methodologists over the last 15 years (e.g., **?**, **?**, **?**, **?**, **?**). Whereas **?** sought to derive qualitative principles of causal inference within a correlational framework, qualitative methodologists writing in the wake of "KKV" have emphasized and clarified process-tracing's "within-case" inferential logic: in process tracing, explanatory hypotheses are tested based on observations of what happened within a case, rather than on covariation between causes and effects across cases. The process tracing literature has also advanced increasingly

[1] As we describe in **?**, the term "qualitative research" means many different things to different scholars, and there are multiple approaches to mixing qualitative and quantitative methods. There we distinguish between approaches that suggest that qualitative and quantitative approaches address distinct, if complementary, questions; those that suggest that they involve distinct measurement strategies; and those that suggest that they employ distinct inferential logics. The approach that we employ in **?** connects most with the third family of approaches. Most closely related, in political science, is work in **?**, in which researchers use knowledge about the empirical joint distribution of the treatment variable, the outcome variable, and a post-treatment variable, alongside assumptions about how causal processes operate, to tighten estimated bounds on causal effects. In the present book, however, we move toward a position in which fundamental differences between qualitative and quantitative inference tend to dissolve, with all inference drawing on what might be considered a "qualitative" logic in which the researcher's task is to confront a pattern of evidence with a theoretical logic.

elaborate conceptualizations of the different kinds of probative value that within-case evidence can yield.

For instance, qualitative methodologists have explicated the logic of different test types ("hoop", "smoking gun", etc.) involving varying degree of specificity and sensitivity (**?**, **?**). A smoking-gun test is a test that seeks information that is only plausibly present if a hypothesis is true (thus, generating strong evidence for the hypothesis if passed), a hoop test seeks data that should certainly be present if a proposition is true (thus generating strong evidence against the hypothesis if failed), and a doubly decisive test is both smoking-gun and hoop (for an expanded typology, see also **?**). Other scholars have expressed the leverage provided by process-tracing evidence in Bayesian terms, moving from a set of discrete test types to a more continuous notion of probative value (**?**, **?**, **?**).[2]

Yet, conceptualizing the different ways in which probative value might operate leaves a fundamental question unanswered: what gives within-case evidence its probative value with respect to causal relations? We believe that, fundamentally, the answer lies in researcher beliefs that lies outside of the analysis in question. We enter a research situation with a model of how the world works, and we use this model to make inferences given observed patterns in the data—while at the same time updating those models based on the data. A key aim of this book is to demonstrate how models can — and, in our view, must — play in drawing case-level causal inferences.

As we will also argue, along with clarifying the logic of qualitative inference, causal models can also enable the systematic integration of qualitative and quantitative forms of evidence. Social scientists are increasingly pursuing mixed-method research designs. It is becoming increasingly common for scholars to pursue research strategies that combine quantitative with qualitative forms of evidence. A typical mixed-methods study includes the estimation of causal effects using data from many cases as well as a detailed examination of the processes taking place in a few. Prominent examples in-

[2]In **?**, we use a fully Bayesian structure to generalize Van Evera's four test types in two ways: first, by allowing the probative values of clues to be continuous; and, second, by allowing for researcher uncertainty (and, in turn, updating) over these values. In the Bayesian formulation, use of process-tracing information is not formally used to conduct tests that are either "passed" or "failed", but rather to update beliefs about different propositions.

clude Lieberman's study of racial and regional dynamics in tax policy (**?**); Swank's analysis of globalization and the welfare state (**?**); and Stokes' study of neoliberal reform in Latin America (**?**). Major recent methodological texts provide intellectual justification of this trend toward mixing, characterizing small-n and large-n analysis as drawing on a single logic of inference and/or as serving complementary functions (King, Keohane, and Verba, 1994; Brady and Collier, 2004). The American Political Science Association now has an organized section devoted in part to the promotion of multi-method investigations, and the emphasis on multiple strategies of inference research is now embedded in guidelines from many research funding agencies (Creswell and Garrett, 2008).

However, while scholars frequently point to the benefits of mixing correlational and process-based inquiry (e.g., **?**, p.~181), and have sometimes mapped out broad strategies of multi-method research design (**?**, **?**), they have rarely provided specific guidance on how the integration of inferential leverage should unfold. In particular, the literature does has not supplied specific principles for aggregating findings—whether mutually reinforcing or contradictory—across different modes of analysis.A small number of exceptions stand out. In the approach suggested by **?**, for instance, available expert (possibly imperfect) knowledge regarding the operative causal mechanisms for a small number of cases can be used to anchor the statistical estimation procedure in a large-N study. **?** propose a Bayesian approach in which qualitative information shapes subjective priors which in turn affect inferences from quantitative data. Relatedly, in **?**, researchers use knowledge about the empirical joint distribution of the treatment variable, the outcome variable, and a post-treatment variable, alongside assumptions about how causal processes operate, to tighten estimated bounds on causal effects. **?** presents an informal framework in which case studies are used to test the assumptions underlying statistical inferences, such as the assumption of no-confounding or the stable-unit treatment value assumption (SUTVA).

Yet we still lack a comprehensive framework that allows us to enter qualitative and quantitative form of information into an integrated analysis for the purposes of answering the wide range of causal questions that are of interests to social scientists, including questions about case-level explanations and causal effects, average causal effects, and causal pathways. As we aim to demonstrate in this book, grounding inference in causal models provides a very natural way of combining information of the X, Y variety with in-

formation about the causal processes connecting X and Y. The approach can be readily addressed to both the case-oriented questions that tend to be of interest to qualitative scholars and the population-oriented questions that tend to motivate quantitative inquiry. As will become clear, in fact, when we structure our inquiry in terms of causal models, the conceptual distinction between qualitative and quantitative inference becomes hard to sustain. Notably, this is not for the reason that "KKV"'s framework suggests, i.e., that all causal inference is fundamentally about correlating causes and effects. To the contrary, it is that in a causal-model-based inference, what matters for the informativeness of a piece of evidence is how that evidence is connected to our query, given how we think the world works. While the apparatus that we present is formal, the approach—in asking how pieces of evidence drawn from different parts of a process map on to a base of theoretical knowledge—is arguably most closely connected to process tracing in its core logic.

1.1.3 Connecting theory and empirics

Theory and empirics have had a surprisingly uncomfortable relationship in political science. In a major recent intervention, for instance, **?** draw attention to and critique political scientists' extremely widespread reliance on the "hypothetico-deductive" (H-D) framework, in which a theory or model is elaborated, empirical predictions derived, and data sought to test these predictions and the model from which they derive.Clarke and Primo draw on decades of scholarship in the philosophy of science pointing to deep problems with the HD framework, including with the idea that the truth of a model logically derived from first principles can be *tested* against evidence.

This book is also motivated by a concern with the relationship between theory and evidence in social inquiry. In particular, we are struck by the frequent lack of a clear link between theory, on the one hand, and empirical strategy and inference, on the other. We see this ambiguity as relatively common in both qualitative and quantitative work. We can perhaps illustrate it best, however, by reference to qualitative work, where the centrality of theory to inference has been most emphasized. In process tracing, theory is what justifies inferences. In their classic text on case study approaches, **?** describe process tracing as the search for evidence of "the causal process that a theory hypothesizes or implies" (6). Similarly, **?** conceptualizes the approach as testing for the causal-process-related observable implications of a theory, **?**

indicates that the events for which process tracers go looking are those posited by theory (128), and **?** describes theory as a source of predictions that the case-study analyst tests (116). Theory, in these accounts, is supposed to help us figure out where to look for discriminating evidence.

What we do not yet have, however, is a systematic account of how researchers can derive within-case empirical predictions from theory and how exactly doing so provides leverage on a causal question. From what elements of a theory can scholars derive informative within-case observations? Given a set of possible things to be observed in a case, how can theory help us distinguish more from less informative observations? Of the many possible observations suggested by a theory, how can we determine which would add probative value to the evidence already at hand? How do the evidentiary requisites for drawing a causal inference, given a theory, depend on the particular causal question of interest—on whether, for instance, we are interested in identifying the cause of an outcome, estimating an average causal effect, or identifying the pathway through which an effect is generated? In short, how exactly can we ground causal inferences from within-case evidence in background knowledge about how the world works?

Most quantitative work in political science features a similarly weak integration between theory and research design. The modal inferential approach in quantitative work, both observational and experimental, involves looking for correlations between causes and outcomes, with minimal regard for intervening or surrounding causal relationships.[3]

In this book, we seek to show how scholars can make much fuller and more explicit use of theoretical knowledge in designing their research projects and analyzing their observations. Like Clarke and Primo, we treat models not as maps of sort: maps, based on prior theoretical knowledge, about causal relations in a domain of interest. Also as in Clarke and Primo's approach, we do not write down a model in order to test its veracity. Rather, we show how we can systematically use causal models with particular characteristics to guide our empirical strategies and inform our inferences. Grounding our empirical strategy in a model allows us, in turn, to learn about the model itself as we encounter the data.

[3]One exception is structural equation modeling, which bears a close affinity to the approach that we present in this book, but has gained minimal traction in political science.

1.2 Key contributions

This book draws on methods developed in the study of Bayesian networks, a field pioneered by scholars in computer science, statistics, and philosophy. Bayesian networks, a form of causal model, have had limited traction to date in political science. Yet the literature on Bayesian networks and their graphical counterparts, directed acyclic graphs (DAGs), is a body of work that addresses very directly the kinds of problems that qualitative and quantitative scholars routinely grapple with.[4]

Drawing on this work, we show in the chapters that follow how a theory can be formalized as a causal model represented by a causal graph and a set of structural equations. Engaging in this modest degree of formalization yields enormous benefits. It allows us, for a wide range of causal questions, to specify causal questions clearly and assess what inferences to make about queries from new data.

For students engaging in process tracing, the payoffs of this approach are that it provides:

- A grounding for assessing the "probative value" for data from different parts of any causal network.
- A way of aggregating inferences from observations drawn from different parts of the causal network in a way that is transparent and replicable.
- Guidance for research design: formalization can be used to assess the relative informativenss of of different evidentiary and case-selection

[4]For application to quantitative analysis strategies in political science, **?** give a clear introduction to how these methods can be used to motivate strategies for conditioning and adjusting for causal inference; **?** demonstrate how these methods can be used to assess claims of external validity. With a focus on qualitative methods, **?** uses causal diagrams to lay out a ''completeness standard" for good process tracing. **?** employ graphs to conceptualize the different possible pathways between causal and outcome variables among which qualitative researchers may want to distinguish. Generally, in discussions of qualitative methodology, graphs are used to capture core features of theoretical accounts, but are not developed specifically to ensure a representation of the kind of independence relations implied by structural causal models (notably what is called in the literature the "Markov condition"). Moreover, efforts to tie these causal graphs to probative observations, as in **?**, are generally limited to identifying steps in a causal chain that the researcher should seek to observe.

strategies, conditional on how you think the world works and the question you want to answer.

For mixed method inference:

- Systematic integration — using both qual and quant to both help answer any given query. in fact, no fundamental difference between quant and qual data — which may discomfit some readers, who see qual research as fundamentally distinct, but offers big advantages, including:
- Transparency: how exactly the qual and the quant enter into the analysis.
- A way to justify the background assumptions you've used
- Learning in both directions: from cases to populations, from populations to cases
- Which provides a model for cumulation. Models get updated and become priors for new analyses.
- Design: diagnosis of wide vs deep, as well as evidentiary and case-selection strategies

As we will show, using causal models has substantial implications for common methodological advice and practice. To touch on just a few of these: Our elaboration and application of model-based process tracing shows that, given plausible causal models, process tracing's common focus on intervening causal chains may be much less productive than other empirical strategies, such as examining moderating conditions. Our examination of model-based case-selection indicates that for many common purposes there is nothing particularly especially informative about "on the regression line" cases or those in which the outcome occurred, and that case selection should often be driven by factors that have to date received little attention, such as the population distribution of cases and the probative value of the available evidence. And an analysis of clue-selection as a decision problem shows that the probative value of a piece evidence cannot be assessed in isolation, but hinges critically on what we have already observed.

The basic analytical apparatus that we employ in this book is not new. Rather, we see the book's goals as being of three kinds. First, the book aims to import insight: to introduce political scientists to an approach that

has received little attention in the discipline but that can be useful for addressing the sorts of causal causal questions with which political scientists are commonly preoccupied. As a model-based approach, it is a framework especially well suited to a field of inquiry in which exogeneity frequently cannot be assumed by design—that is, in which we often have no choice but to be engineers. Second, the book draws connections between the Bayesian networks approach and key concerns and challenges with which students in social sciences routinely grapple. Working with causal models and DAGs most naturally connects to concerns about confounding and identification that have been central to much quantitative methodological development. Yet we also show how causal models can address issues central to process tracing, such as how to select cases for examination, how to think about the probative value of causal process observations, and how to structure our search for evidence, given finite resources. Third, the book provides a set of usable tools for implementing the approach. We provide intuition and software that researchers can use to make research design choices and draw inferences from the data.

1.3 The Road Ahead

The book is divided into four main parts.

The first part is about the basics. We start off by describing the kinds of causal estimands of interest. The main goal here is to introduce the key ideas in the study of Bayesian nets and to argue for a focus of interest away from average treatment effects as go-to estimands of interest and towards a focus on causal nets, or causal structures, as the key quantity of interest. The next chapter introduces key Bayesian ideas; what Bayes' rule is and how to use it. The third chapter connects the study of Bayesian networks to theoretical claims. The key argument here is that nets should be thought of as theories which are themselves supportable by lower level networks (theories). Lower level theories are useful insofar as they provide leverage to learn about processes on higher level networks.

The second part applies these ideas to process tracing and mixed methods designs. Rather than conceptualizing process tracing as has been done in recent work as seeking process level data that is known to be informative about a causal claim, the approach suggested here is one in which the probative

value of a clue is derived from its position in a causal network connecting variables of interest. Chapter 5 lays out the key logic of inference from clues and provides general criteria for assessing when it is and is not possible. Chapter 6 provides specific tools for assessing which collections of clues are most informative for a given estimand of interest and outlines a strategy for assessing which clues to gather when in a research process. Chapter 7 applies these tools to the problem of assessing the effects of economic inequality on democratization.

Chapter 8 moves to mixed data problems — situations in which a researcher contains "quantitative" (X, Y) data on a set of cases and is considering gathering within case ("qualitative") data on some of these. We argue that this situation is formally no different to the single case process tracing problem since a collection of cases can always be conceptualized as a single case with vector valued variables. The computational complexity is however greater in these cases and so in this chapter we describe a set of models that may be useful for addressing these problems. We conclude this part by revisiting the problem of inequality and democracy introduced in Chapter 7.

The third part focuses on research design. In this framework the problem of case selection is equivalent to the kind of problem of clue selection discussed in Chapter 6. For a canonical multicase model however we use simulation approaches to provide guidance for how cases should be selected. The broad conclusion here is that researchers should go where the probative value lies, and all else equal, should select cases approximately proportional to the size of XY strata—whether or not these are "on the regression line."

The fourth part steps back and puts the model-based approach into question. We have been advocating an embrace of models to aid inference. But the dangers of doing this are demonstrably large. The key problem is that with model-based inference, the inferences are only as good as the model. In the end, while we are supporting the efforts of engineers, we know that the philosopher is right. This final part provides four responses to this (serious) concern. The first is that the dependence on models can sound more extreme than it is. Seemingly fixed parameters of models can themselves become quantities of interest in lower-level models, and there can be learning about these when higher-level models are studied. Thus models are both put to use and objects of interest. The second is that different types of conditional statements are possible; in particular as shown in work qualitative

graphs. The third response points to the sort of arguments that can be made to support models, most importantly the importation of knowledge from one study to another. The last argument, presented in the last substantive chapter, highlights the tools to *evaluate* models, using approaches that are increasingly standard in Bayesian analysis.

Here we go.

Part I

Foundations

Chapter 2

Causal Models

We provide a lay language primer on the logic of causal models.

Causal knowledge is not just the end goal of much empirical social science; it is also often a key input into causal inference. Rarely do we arrive at causal inquiry fully agnostic about causal relations in the domain of interest. As nicely put by **?**, no causes in, no causes out. Moreover, our beliefs about how the world works—as we show later in this book—have profound implications for how the research process and inference should unfold.

What we need is a language for expressing our prior causal knowledge such that we can full exploit it, drawing inferences and making research design decisions in ways that are logically consistent with our beliefs, and such that others can readily see and assess those underlying premises. Causal models provide such a language.

In this chapter we provide a basic introduction to causal models. Subsequent chapters in Part I layer on other foundational components of the book's framework, including a causal-model-based understanding of theory, the definition of common causal estimands within causal models, and the basics of Bayesian inference. While here we focus on the formal definition of causal models, in Chapter 10 we discuss strategies for generating them.

2.1 The counterfactual model

We begin with what we might think of as a meta-model, the counterfactual model of causation. The counterfactual model is the dominant approach to causal relations in the social sciences. At its core, a counterfactual understanding of causation captures a simple notion of causation as "difference-making."[1] In the counterfactual view, to say that X caused Y is to say: *had* X been different, Y *would have been* different. Critically, the antecedent, "had X been different," imagines a *controlled* change in X—an intervention that altered X's value—rather than a naturally arising difference in X. The counterfactual claim, then, is not that Y is different in those cases in which X is different; it is, rather, that if one could have *made* X different, Y would have been different.

Turning to a substantive example, consider, for instance, the claim that India democratized (Y) because it had a relatively high level of economic equality (X) (drawing on the logic of **?**). In the counterfactual view, this is equivalent to saying that, had India *not* had a high level of equality—where we imagine that we *made* equality in India lower—the country would not have democratized. High economic equality made a difference.

Along with this notion of causation as difference-making, we also want to allow for variability in how X acts on the world. X might sometimes make a difference, for some units of interest, yet sometimes not. High levels of equality might generate democratization in some countries or historical settings but not in others. Moreover, while equality might make democratization happen in some times in places, it might prevent that same outcome in others. In political science, we commonly employ the "potential outcomes" framework to describe the different kinds of counterfactual causal relations that might prevail between variables (**?**). In this framework we characterize how a given unit responds to a causal variable by positing the outcomes that it *would* take on at different values of the causal variable.

It is quite natural to think about potential outcomes in the context of medical treatment. Consider a situation in which some individuals in a diseased

[1]The approach is sometimes attributed to David Hume, whose writing contains ideas both about causality as regularity and causality as counterfactual. On the latter the key idea is "if the first object had not been, the second never had existed" (**?**, Section VIII). More recently, the counterfactual view has been set forth by **?** and **?**. See also **?**.

population are observed to have received a drug ($X = 1$) while others have not ($X = 0$). Assume that, subsequently, a researcher observes which individuals become healthy ($Y = 1$) and which do not ($Y = 0$). Let us further stipulate that each individual belongs to one of four unobserved response ''types," defined by the potential effect of treatment on the individual:[2]

- **a**dverse: Those who would get better if and only if they do not receive the treatment
- **b**eneficial: Those who would get better if and only if they do receive the treatment
- **c**hronic: Those who will remain sick whether or not they receive treatment
- **d**estined: Those who will get better whether or not they receive treatment

We can express this same idea by specifying the set of "potential outcomes" associated with each type of patient, as illustrated in Table **??**.

Table 2.1: . Potential outcomes: What would happen to each of four possible types of case if they were or were not treated.

	Type a	Type b	Type c	Type d
	adverse effects	**b**eneficial Effects	**c**hronic cases	**d**estined cases
Not treated	Healthy	Sick	Sick	Healthy
Treated	Sick	Healthy	Sick	Healthy

In each column, we have simply written down the outcome that a patient of a given type would experience if they are not treated, and the outcome they would experience if they are treated.

Throughout the book, we generalize from this toy example. Whenever we have one causal variable and one outcome, and both variables are binary (i.e., each can take on two possible values, 0 or 1), then there are only four sets of possible potential outcomes, or "causal types." More generally, for any

[2]We implicitly invoke the assumption that the treatment or non-treatment of one patient has no effect on the outcomes of other patients. This is known as the stable unit treatment value assumption (SUTVA). See also **?** for a similar classification of types.

pair of causal and outcome variables, we will use θ^Y to denote the causal type at node Y. We, further, add subscripts to denote particular types, as for instance with θ^Y_{ij}. Here i represents the case's potential outcome when $X = 0$ and j is the case's potential outcome when $X = 1$.

Incorporating this notation, when we have one binary causal variable and a binary outcome, the four types are:

- **a**: A negative causal effect of X on Y. We write this as: $\theta^Y = \theta^Y_{10}$.
- **b**: A positive causal effect of X on Y. We write this as: $\theta^Y = \theta^Y_{01}$.
- **c**: No causal effect, with Y "stuck" at 0. We write this as: $\theta^Y = \theta^Y_{00}$.
- **d**: No causal effect, with Y "stuck" at 1. We write this as: $\theta^Y = \theta^Y_{11}$.

Table **??** summarizes these types in terms of potential outcomes:

Table 2.2: . Generalizing from Table **??**, the table gives for each causal type the values that Y would take on if X is set at 0 and if X is set at 1.

	Type a	Type b	Type c	Type d
	$\theta^Y = \theta^Y_{10}$	$\theta^Y = \theta^Y_{01}$	$\theta^Y = \theta^Y_{00}$	$\theta^Y = \theta^Y_{11}$
Set $X = 0$	$Y(0) = 1$	$Y(0) = 0$	$Y(0) = 0$	$Y(0) = 1$
Set $X = 1$	$Y(1) = 0$	$Y(1) = 1$	$Y(1) = 0$	$Y(1) = 1$

Returning to our democratization example, let $I = 1$ represent a high level of economic equality and $I = 0$ its absence, with $D = 1$ representing democratization and $D = 0$ its absence. A θ^D_{10} (a) type, then, is any case in which a high level of equality, if it occurs, *prevents* democratization in a country that would otherwise have democratized. The causal effect of high equality in an a type is $= -1$. A θ^D_{01} (b) type is a case in which high equality, if it occurs, generates democratization in a country that would otherwise have remained non-democratic (effect = 1). A θ^D_{00} (c) type is a case that will not democratize regardless of the level of equality (effect = 0); and a θ^D_{11} (d) type is one that will democratize regardless of the level of equality (again, effect = 0).

In this setting, a causal *explanation* of a given case outcome amounts to

a statement about its type. The claim that India democratized because of a high level of equality is equivalent to saying that India democratized and is θ^D_{01} type. To claim that Sierra Leone democratized because of low inequality is equivalent to saying that Sierra Leone democratized and is θ^D_{10} type. To claim, on the other hand, that Malawi democratized for reasons having nothing to do with its level of economic equality is to characterize Malawi as a θ^D_{11} type (which already specifies its outcome).

2.1.1 Generalizing to outcomes with many causes

We can also use potential-outcomes reasoning for more complex causal relations. For example, supposing there are two binary causal variables X_1 and X_2, we can specify any given case's potential outcomes for each of the different possible combinations of causal conditions—there now being four such conditions (as each causal variable may take on 0 or 1 when the other is at 0 or 1).

As for notation, we now need to expand θ's subscript since we need to represent the value that Y takes on under each of the four possible combinations of X_1 and X_2 values. We construct the four-digit subscript to with the ordering:

$$Y_{hijk}\begin{cases} h &= Y|(X_1=0, X_2=0) \\ i &= Y|(X_1=1, X_2=0) \\ j &= Y|(X_1=0, X_2=1) \\ k &= Y|(X_1=1, X_2=1) \end{cases}$$

Thus, for instance, θ^Y_{0100} means that Y is 1 if $X_1 = 1$ and $X_2 = 0$ and is 0 otherwise. We now have 16 causal types: 16 different patterns that Y might display in response to changes in X_1 and X_2. The full set is represented in Table **??**, which also makes clear how types are read off of four-digit subscripts. (The type numberings in the first column are, of course, arbitrary here and included for ease of reference.)

We can read off this table that for nodal type θ^Y_{0101}, X_1 has a positive causal effect on Y but X_2 has no effect, whereas for θ^Y_{0011}, X_2 has a positive effect but X_1 has none. We also capture interactions here. For instance, θ^Y_{0001}, X_2 has a positive causal effect if and only if X_1 is 1. In that case X_1 and X_2

Table 2.3: With two binary causal variables, there are 16 causal types: 16 ways in which Y might respond to changes in the two variables.

θ^Y	if $X_1=0, X_2=0$	if $X_1=1,X_2=0$	if $X_1=0,X_2=1$	if
θ^Y_{0000}	0	0	0	
θ^Y_{1000}	1	0	0	
θ^Y_{0100}	0	1	0	
θ^Y_{1100}	1	1	0	
θ^Y_{0010}	0	0	1	
θ^Y_{1010}	1	0	1	
θ^Y_{0110}	0	1	1	
θ^Y_{1110}	1	1	1	
θ^Y_{0001}	0	0	0	
θ^Y_{1001}	1	0	0	
θ^Y_{0101}	0	1	0	
θ^Y_{1101}	1	1	0	
θ^Y_{0011}	0	0	1	
θ^Y_{1011}	1	0	1	
θ^Y_{0111}	0	1	1	
θ^Y_{1111}	1	1	1	

are "complements." For θ^Y_{0111}, X_2 has a positive causal effect if and only if X_1 is 0. In that case X_1 and X_2 are "substitutes."

As one might imagine, the number of causal types increases rapidly (very rapidly) as the number of considered causal variables increases, as it also would if we allowed X or Y to take on more than 2 possible values. For example if there are n binary causes of an outcome then there can be $2^{(2^n)}$ causal types of this form. However, the basic principle of representing possible causal relations as patterns of potential outcomes remains unchanged, at least as long as variables are discrete.

A somewhat counter-intuitive implication of the counterfactual framework lies in how it forces us to think about multiple causes. When seeking to explain the outcome in a case, researchers sometimes proceed as though competing explanations amount to *rival* causes, where X_1 being a cause of Y implies that X_2 was not. Did Malawi democratize because it was a relatively

economically equal society *or* because of international pressure to do so? In the counterfactual model, however, causal relations are non-rival. If two out of three people vote for an outcome under majority rule, for example, then both of the two supporters caused the outcome: the outcome would not have occurred if *either* supporter's vote were different. Put differently, when we say that X caused Y in a given case, we will generally mean that X was *a* cause, X will rarely be *the* cause in the sense of being the *only* thing a change in which would have changed the outcome. Malawi might not have democratized if *either* a relatively high level of economic equality or international pressure had been absent. For most social phenomena that we study, there will be multiple, and sometimes a great many, difference-makers for any given case outcome.

2.1.2 Deterministic relations

You might notice that in the counterfactual framework, as we have described it, causal relations are conceptualized as deterministic. A given case has a set of potential outcomes. If we know the type, any uncertainty about outcomes enters as incomplete knowledge of the factors influencing an outcome. But, in principle, if we knew all of the relevant causal conditions and the complete set of potential outcomes for a case, we could perfectly predict the actual outcome in that case. This understanding of causality—as ontologically deterministic, but empirically imperfectly understood—is compatible with views of causation commonly employed by qualitative researchers (see, e.g., **?**), and with understandings of causal determinism going back at least to **?**. As we will see, we can readily express this kind of incompleteness of knowledge within a causal model framework; indeed, the way in which causal models manage uncertainty is central to how they allow us to pose questions of interest and to learn from evidence.

2.2 Causal Models and Directed Acyclic Graphs

Potential outcomes tables can capture quite a lot. We could, for instance, summarize our beliefs about the relationship between economic equality and democratization by saying that we think that the world is comprised of a mixture of a, b, c, and d types, as defined above. We could get more specific

and express a belief about what proportions of cases in the world are of each of the four types. For instance, we might believe that a types and d types are quite rare while b and c types are more common. Moreover, our belief about the proportions of b (positive effect) and a (negative effect) cases imply a belief about equality's *average* effect on democratization as, in a binary setup, this quantity is simply the proportion of b types minus the proportion of a types.

As we have seen, beliefs about even more complex causal relations can be fully expressed in potential-outcomes notation. However, as causal structures become more complex—especially, as the number of variables in a domain increases—a causal model can be a powerful organizing tool. In this section, we show how causal models and their visual counterparts, directed acyclic graphs (DAGs), can represent substantive beliefs about counterfactual causal relationships in the world. The key ideas in this section can be found in many texts (see, e.g., Halpern and Pearl (2005) and Galles and Pearl (1998)), and we introduce here a set of basic principles that readers will need to follow the argumentation in this book.

To slightly shift the frame of our running example, suppose that we believe the level of economic inequality can have an effect on whether a country democratizes. We might believe inequality affects the likelihood of democratization by generating demands for redistribution, which in turn can cause the mobilization of lower-income citizens, which in turn can cause democratization. We might also believe that mobilization itself is not just a function of redistributive preferences but also of the degree of ethnic homogeneity, which shapes capacities of lower-income citizens for collective action. We can visualize this model in Figure **??**.

2.2.1 Components of a Causal Model

In the context of this example, let us now consider the three components of a causal model: variables, functions, and distributions.

2.2.1.1 The variables.

The first component of a causal model is the set of variables across which the model characterizes causal relations. On the graph in Figure **??**, the 6 included variables are represented by the 6 nodes.

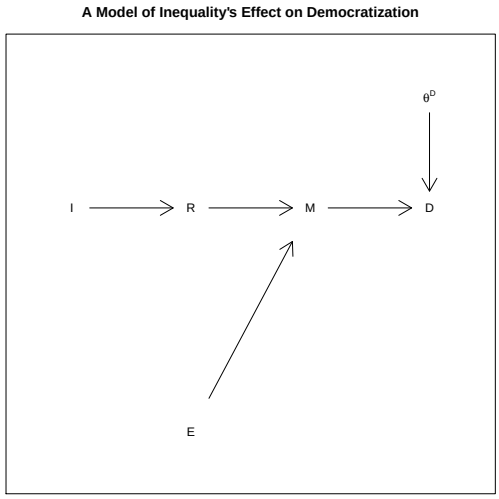

Figure 2.1: A simple causal model in which high inequality (I) affects the democratization (D) via redistributive demands and mass mobilization (M), which is also a function of ethnic homogeneity (E). The arrows show relations of causal dependence between variables. The graph does not capture the ranges of the variables and the functional relations between them.

In a causal-model framework, we sometimes use familial terms to describe relations among variables. For instance, two nodes directly connected by an arrow are known as "parent" and "child," while two nodes with a child in common (both directly affect the same variable) are "spouses." We can also say that I is an "ancestor" of D (a node upstream from D's parent) and conversely that D is a descendant of I (a node downstream from I's child).

In identifying the variables, we also need to specify the *ranges* across which they can potentially vary. We might specify, for instance, that all variables in the model are binary, taking on the values 0 or 1. We could, alternatively, define a set of categories across which a variable ranges or allow a variable to take on any real number value or any value between a set of bounds. [3]

Notice that some of these variables have arrows pointing *into* them: R, M, and D are endogenous variables, meaning that their values are determined entirely by other variables in the model.

Other nodes have arrows pointing out of them but no arrows pointing into them: I, E and θ^D. I and E are "exogenous" nodes, they influence other variables in the model but themselves have no causes specified in the model.

θ^D requires a little more explanation since it does not describe a substantive variable. In the world of causal models, U terms are typically used to capture unspecified exogenous influences. We could have, here, included a term U_D to indicate an "error" term or uncertainty regarding exactly what value D will take given knowledge of I and E. We have used θ^D however to highlight the fact in non parametric models this "residual" component can be thought of as *the* locus of learning about the questions we are asking. θ^D can be thought of as capturing *how* the parents of D produce D. In the present example, we believe democratization to be potentially affected by mobilization, but we also know that democratization is affected by other things, even if we do not know what they are. We can thus think of θ^D (equivalently) as capturing a set of unknown factors—factors other than mobilization—that affect democratization and the ways known factors produce the outcome.

[3] If we let $\mathcal{R}$ denote a set of ranges for all variables in the model, we can indicate X's range, for instance, by writing $\mathcal{R}(X) = \{0, 1\}$. The variables in a causal model together with their ranges—the triple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$—are sometimes called a *signature*, $\mathcal{S}$.

2.2.1.2 The functions.

Next, we need to specify our beliefs about the causal relations among the variables in our model. How is the value of one variable affected by, and how does it affect, the values of others? For each endogenous variable—each variable influenced by others in the model—we need to express beliefs about how its value is affected by its parents, its immediate causes.

The graph already represents some aspects of these beliefs: the arrows, or directed edges, tell us which variables we believe to be direct causal inputs into other variables. So, for instance, we believe that democratization (D) is determined jointly by mobilization (M) and some exogenous, unspecified factor (or set of factors), θ^D. We can think of θ^D as all of the other influences on democratization, besides mobilization, that we either do not know of or have decided not to explicitly include in the model. We believe, likewise, that M is determined by I and an unspecified exogenous factor (or set of factors), θ^M. And we are conceptualizing inequality (I) as shaped solely by a factors exogenous to the model, captured by θ^I. (For all intents and purposes, I behaves as an exogenous variable here since its value is determined solely by an exogenous variable.)

We can also, however, express more specific beliefs about causal relations in the form of a causal function.[4] Specifying a function means writing down whatever general or theoretical knowledge we have about the direct causal relations between variables. A function specifies how the value that one variable takes on is determined by the values that other variables—its parents—take on.

We can specify this relationship in a vast variety of ways. It is useful however to distinguish broadly between parametric and non parametric approaches.

- A *parametric* approach specifies a functional form that relates parents to children. For instance we might model one variable as a linear function of another. For instance, we can write $R = \beta I$, where β is a parameter that we do not know the value of at the outset of a study but which we wish to learn about. If we believe D to be linearly affected by M but also subject to forces that we do not yet understand and have not yet specified in our theory, then we can write: $D = \beta M + U_D$, where U_D represents a random disturbance. We can be still more agnostic

[4] The collection of all causal functions in the model can be denoted as $\mathcal{F}$.

by, for example including parameters that govern how other parameters operate. Consider, for instance the function, $D = \beta M^{U_D}$. Here, D and M are linearly related if $U_D = 1$, but exponentially if U_D is anything other than 1. The larger point is that functions can be written to be quite specific or extremely general, depending on the state of prior knowledge about the phenomenon under investigation. The use of a structural model *does not require precise knowledge of specific causal relations*, even of the functional forms through which two variables are related.

- With discrete data, causal functions can also take fully *non-parametric* form, allowing for *any possible relation* between parents and children. Let us, for instance, allow U_D to range across the four possible values, yielding the following causal function for D:
  - if $U_D = \theta^D_{10}$, then $D = 1 - M$
  - if $U_D = \theta^D_{01}$, then $D = M$
  - if $U_D = \theta^D_{00}$, then $D = 0$
  - if $U_D = \theta^D_{11}$, then $D = 1$

We are, of course, drawing on our original four causal types from earlier in this chapter. Here, U_D is essentially a placeholder for causal types. We can think of it as an unknown factor that conditions the effect of mobilization on democratization, determining whether M has a negative effect, a positive effect, no effect with democratization never occurring, or no effect with democratization bound to occur regardless of mobilization.

Using our causal type framework, we can similarly use U terms to designate causal relations involving of any number of parent nodes. With two parent nodes, for instance, we simply use causal types of the form θ^Y_{hijk}, as illustrated above.

The chapters to come operate in a non-parametric vein, with U terms as receptacles for causal types. To emphasize this feature, we continue as we do in Figure **??** to use θ instead of U to represent case specific features. Thus, every substantively defined node, J, in a graph has a θ^J term pointing into it, and the value of θ^J gives the mapping from J's parents (if it has any) to the value of J. The basic idea, applied to the binary variables in Figure **??**, is as follows:

- **Nodes with no parents**: For an exogenous node like I, θ^I represents

an "assignment" process and can take on one of two values, θ^I_0, meaning that I is "assigned" to 0 or θ^I_1, meaning that I is assigned to 1.

- **Binary nodes with 1 binary parent**: For endogenous node R, with only one parent (I), θ^R takes on one of four values of the form θ^R_{ij} (our four original types, θ^R_{10}, θ^R_{01}, etc.).
- **Binary nodes with 2 binary parents**: θ^M will take on a possible 16 values of the form θ^M_{hijk} (θ^M_{0000}, θ^M_{0001}, etc.).
- **Binary nodes with n binary parents** have $2^{(2^n)}$ subscripts.

For analytic applications later in the book, we will want to be able to think both about the causal type operation at a particular *node* and about *collections* of causal types across a model. We thus refer to θ^J as a unit's *nodal* causal type, or simply nodal type, for J.[5] We refer to the collection of nodal types across all nodes for a given unit (i.e., a case) as the case's *unit causal type*, or simply *causal type*, denoted by θ. Since the nodal types of exogenous include values of exogenous nodes, then the unit's causal type fully specifies all node values as well as all *counterfactual* node values for a unit.

Box: Nodal types, causal types

term	symbol	meaning
nodal type	θ^X	The way that a node responds to the values of its parents. Example: θ^Y_{10}, written `Y10`: Y takes the value 1 if $X = 0$ and 0 if $X = 1$.
causal type	θ	A causal type is a concatenation of nodal types, one for each node. Example: $(\theta^X_0, \theta^Y_{00})$, written `X0.Y00`, is a type that has $X = 0$ and $Y = 0$ no matter what the value of X.

---

[5] The types here map directly into the four types, a, b, c, d, used in **?** and into principal strata employed by Rubin and others. The literature on probabilistic models also refers to such strata as "canonical partitions" or "equivalence classes." Note that this model is not completely general as the multinomial distribution assumes that errors are iid.

It is thus worth dwelling for a moment on what this kind of function is doing. We have started with a graph in which mobilization can have an effect on democratization and the understanding that this effect, both its existence and its sign, may vary across cases. Cases, in other words, may be of different causal types. Further, we do not know what it is that shapes D's response to M—what makes a case one type versus another. We thus use θ_D as a stand-in for the unknown and unspecified moderators of M's effect. We might, at this stage, wonder what the point is of including θ_D in the model; are we not essentially just placing a question mark on the graph? We are, and that is precisely the point. As we will see in later chapters, non-substantive, causal-type nodes can play a key role in specifying (a) what we are uncertain about in a causal network and (b) what we would like to find out. Embedding our questions about the world directly into a model of the world, in turn, allows us to answer those questions in ways systematically and transparently guided by prior knowledge.

A few important aspects of causal functions stand out. First, these functions express *causal* beliefs. When we write $D = \beta M$ as a function, we do not just mean that we believe the values of M and D in the world to be linearly related. We mean that we believe that the value of M *determines* the value of D through this linear function. Functions are, in this sense, meant as *directional* statements, with causes on the righthand side and an outcome on the left.

Second, to specify functions is to unpack a potentially complex web of causal relations into its constituent causal links. For each variable, we do not need to think through entire sequences of causation that might precede it. We need only specify how we believe it to be affected by its parents—that is to say, those variables pointing directly into it. Our outcome of interest, D, may be a shaped by multiple, long chains of causality. To theorize how D is generated, however, we write down how we believe D is shaped by its immediate causes, M and θ^D. We then, separately, express a belief about how M is shaped by *its* direct causes, R and E. A variable's function must include as inputs all, and only, those variables that point directly into that variable.[6]

[6]The set of a variable's parents is required to be minimal in the sense that a variable is not included among the parents if, given the other parents, the child does not depend on it in any state that arises with positive probability.

Third, as in the general potential-outcomes framework, all relations in a causal model are conceptualized as in principle deterministic. There is not as much at stake here though as you might think at first; by this we simply mean that a variable's value is *determined* by the values of its parents *along with* any stochastic or unknown components. We express uncertainty about causal relations, however, either as unknown parameters (e.g., β, above) or as random disturbances, the U terms, or the causal types θ.

Fourth, in a properly specified causal model *the values of the exogenous variables*—those with no arrows pointing in to them—*are sufficient to determine the values of all other variables in the model.* Consistent with more informal usage, we refer to a given set of values for all exogenous terms in a model as a *context.* In causal model, context determines all other values. For instance, in Figure **??**, knowing the values of I, E, and θ^D as well as the causal functions (including the values of any parameters they contain) would tell us the values of R, M, and D.

2.2.1.3 The distributions

Putting these components together gives what is termed a *structural causal model.* In a structural causal model, all endogenous variables are, either directly or by implication, functions of a case's context (the values of the set of exogenous variables).[7] What we have not yet inscribed into the model, however, is any beliefs about how *likely* or *common* different kinds of contexts might be. Thus, for instance, a structural causal model consistent with Figure **??** stipulates that I, E, and θ^D may have effects on D, but it says nothing in itself about the distribution of I, E, and θ^D themselves, beyond limitations on their ranges.[8] We have not said anything, for instance, about how common high inequality is across the relevant domain of cases, how common ethnic homogeneity is, or how unspecified inputs are distributed.

[7]More formally, a **structural causal model** *over* signature $\mathcal{S} =< \mathcal{U}, \mathcal{V}, \mathcal{R} >$ is a pair $< \mathcal{S}, \mathcal{F} >$, where $\mathcal{F}$ is a set of ordered structural equations containing a function f_i for each element $Y \in \mathcal{V}$. We say that $\mathcal{F}$ is a set of ordered structural equations if no variable is its own descendant and if no element in $\mathcal{U}$ is parent to more than one element of $\mathcal{V}$. This last condition can be achieved by shifting any parent of multiple children in $\mathcal{U}$ to $\mathcal{V}$. This definition thus includes an assumption of acyclicity, which is not found in all definitions in the literature.

[8]Thus $P(d|i, e, u_D)$ would defined by this structural model (as a degenerate distribution), but $P(i)$, $P(e)$, $P(u_D)$, and $P(i, e, u_D)$ would not be.

In many research situations, we will have beliefs not just about how the world works under different conditions, but also about what kinds of conditions are more likely than others. We can express these beliefs about context as probability distributions over the models exogenous terms.[9] For instance, a structural causal model might support a claim of the form: "R has a positive effect on M if and only if $E = 1$ holds." We might, then, add to this a belief that $E = 1$ in 25% of cases in the population of interest. Including this belief about context implies, in turn, that R has a positive effect on M a quarter of the time. As with the functions, we can also (and typically would) build uncertainty into this belief by specifying a *distribution* over possible shares of cases with ethnic homogeneity, with our degree of uncertainty captured by the distribution's variance.

With our non parametric representation of functional forms, we let λ_j^X denote the probability that $\theta^X = \theta_j^X$. For instance in a simple $X \rightarrow Y$ model, λ_{01}^Y denotes the probability that $\theta^Y = \theta_{01}^Y$.

Technical Note on the Markov Property

The assumptions that no variable is its own descendant and that the U terms are generated independently make the model *Markovian*, and the parents of a given variable are Markovian parents. Knowing the set of Markovian parents allows one to write relatively simple factorizations of a joint probability distribution, exploiting the fact ("the Markov condition") that all nodes are *conditionally independent* of their nondescendants, conditional on their parents. Variables A and B are "conditionally independent" given C if $P(a|b,c) = P(a|c)$ for all values of a, b and c.
To see how this Markovian property allows for simple factorization of P for Figure **??**, note that $P(X, R, Y)$ can always be written as:

$$P(X, R, Y) = P(X)P(R|X)P(Y|R, X)$$

[9] We assume that the exogenous terms, the elements of $\mathcal{U}$, are generated independently of one another. While this is not without loss of generality, it is not as constraining as it might at first appear: any graph in which two exogenous variables are not independent can be replaced by a graph in which these two terms are listed as endogenous (possibly unobserved) nodes, themselves generated by a third variable. Note also that one could envision "incomplete probabilistic causal models" in which researchers claim knowledge regarding distributions over *subsets* of $\mathcal{U}$.

If we believe, as in the figure, that X causes Y only through R then we have the slightly simpler factorization:

$$P(X, R, Y) = P(X)P(R|X)P(Y|R)$$

Or, more generally:

$$P(v_1, v_2, ... v_n) = \prod P(v_i|pa_i) \tag{2.1}$$

The distribution P on $\mathcal{U}$ induces a joint probability distribution on $\mathcal{V}$ that captures not just information about how likely different states are to arise but also the relations of conditional independence between variables that are implied by the underlying causal process. For example, if we thought that X caused Y via R (and only via R), we would then hold that $P(Y|R) = P(Y|X, R)$: in other words if X matters for Y only via R then, conditional on R, X should not be informative about Y.
In this way, a probability distribution P over a set of variables can be consistent with some causal models but not others. This does not, however, mean that a specific causal model can be extracted from P. To demonstrate with a simple example for two variables, any probability distribution on (X, Y) with $P(x) \neq P(x|y)$ is consistent both with a model in which X is a parent of Y and with a model in which Y is a parent of X.

Once we introduce beliefs about the distribution of values of the exogenous terms in a model, we have specified a *probabilistic causal model.* We need not say much more, for the moment, about the probabilistic components of causal models. But to foreshadow the argument to come, our prior beliefs about the likelihoods of different contexts play a central role in the framework that we present in this book. We will see how the encoding contextual knowledge—beliefs that some kinds of conditions are more common than others—forms a key foundation for causal inference. At the same time, our expressions of *uncertainty* about context represent scope for learning: it is the very things that we are, at a study's outset, uncertain about that we can update our beliefs about as we encounter evidence.

2.2.2 Rules for graphing causal models

The diagram in Figure **??** is a causal DAG (**?**). We endow it with the interpretation that an arrow from a parent to a child that a change in the parent can, under some circumstances, induce a change in the child. Though we have already been making use of this causal graph to help us visualize elements of a causal model, we now explicitly point out a number of general features of causal graphs as we will be using them throughout this book. Causal graphs have their own distinctive "grammar," a set of rules that give them important analytic features.

Directed, acyclic. A causal graph represents elements of a causal model as a set of nodes (or vertices), representing variables, connected by a collection of single-headed arrows (or directed edges). We draw an arrow from node A to node B if and only if we believe that A can have a direct effect on B. The resulting diagram is a *directed acyclic* graph (DAG) if there are no paths along directed edges that lead from any node back to itself—i.e., if the graph contains no causal cycles. The absence of cycles (or "feedback loops") is less constraining than it might appear at first. In particular if one thinks that A today causes B tomorrow which in turn causes A today, we can represent this as $A_1 \rightarrow B \rightarrow A_2$ rather than $A \leftrightarrow B$. That is, we timestamp the nodes, turning what might informally appear as feedback into a non cyclical chain.

Meaning of missing arrows. The *absence* of an arrow between A and B means that A is not a direct cause of B.[10] Here lies an important asymmetry: drawing such an arrow does not mean that we know that A *does* directly cause B; but omitting such an arrow implies that we know that A does *not* directly cause B. We say more, in other words, with the arrows we omit than with the arrows that we include.

Returning to Figure **??**, we have here expressed the belief that redistributive preferences exert no direct effect on democratization; we have done so by *not* drawing an arrow directly from R to D. In the context of this model, saying that redistributive preferences have no direct effect on democratization is to say that any effect of redistributive preferences on democratization *must* run through mobilization; there is no other pathway through which such an effect can operate. This might be a way of encoding the knowledge that mass

[10] By "direct" we mean that the A is a parent of B: i.e., the effect of A on B is not fully mediated by one or more other variables in the model.

preferences for redistribution cannot induce autocratic elites to liberalize the regime absent collective action in pursuit of those preferences.

The same goes for the effects of I on M, I on D, and E on D: the graph in Figure **??** implies that we believe that these effects also do not operate directly, but only along the indicated, mediated paths.

Sometimes causes. The existence of an arrow from A to B does not imply that A always has a direct effect on B. Consider, for instance, the arrow running from R to M. The existence of this arrow requires that M appears somewhere in R's functional equation, as a variable's functional equation must include all variables pointing directly into it. Imagine, though, that M's causal function is specified as: $M = RE$. This function allows for the *possibility* that R affects M, as it will whenever $E = 1$. However, it also allows that R will have no effect, as it will when $E = 0$.

This example also, incidentally, demonstrates another important consequence of context, the values of the exogenous variables: a case's context determines not just the settings on the endogenous variables, but also the causal *effects* that prevail among the variables. Under the functional equation $M = RE$, a case's ethnic-compositional context determines whether or not redistributive preferences will have an effect on mobilization.

Representing U on the graph. As a matter of convention, explicitly including U terms is optional. In practice, U's are often excluded from the visual representation of a model on the understanding that every variable on the graph is subject to some unaccounted-for influence and thus, implicitly, has a U term pointing into it. In this book, we will generally draw the U terms where they are of particular theoretical or analytic interest but will otherwise omit them. Whether we include or omit U terms, we will generally treat those nodes in a graph that have no arrows pointing into them as the exogenous variables that define the context.

No excluded common causes, no unobserved confounding Any cause common to multiple variables on the graph must itself be represented on the graph. If A and B on a graph are both affected by some third variable, C, then we must represent this common cause. Put differently, any two variables without common causes on the graph are taken to be independent of one another. Thus, the graph in Figure **??** implies that the values of I, E, and θ^D are all determined independently of one another. If in fact we

believed that a country's level of inequality and its ethnic composition were both shaped by, say, its colonial heritage, then this DAG would *not* be an accurate representation of our beliefs about the world. To make it accurate, we would need to add to the graph a variable capturing that colonial heritage and include arrows running from colonial heritage to both I and E.

This rule ensures that the graph captures all potential correlations among variables that are implied by our beliefs. If I and E are in fact driven by some common cause, then this means not just that these two variables will be correlated but also that each will be correlated with any consequences of the other. For instance, a common cause of I and E would also imply a correlation between R and E. R and E are implied to be independent in the current graph but would be implied to be correlated if a common node pointed into both I and E.

Of particular interest in Figure **??** is the implied independence of θ^D from every other node. Imagine, for instance, an additional node pointing into both I and θ^D. This would represent a classic form of confounding: the assignment of cases to values on the explanatory variable would be correlated with case's potential outcomes on D. The omission of any such pathway is precisely equivalent to expressing the belief that I is exogenous, or (as if) randomly assigned.

Representing excluded common causes, unobserved confounding if you have it. It may be however that there are common causes for nodes that we simply do not understand. We are open to the idea that some unknown feature determines both I and D. In this case it is as if θ^I and θ^D are not independently distributed. This is often represented by adding a dotted line, or a two headed arrow, connecting nodes whose shocks are not independent. Figure **??** illustrates. In general we will allow for this kind of unobserved confounding in the models in this book and seek to learn about the joint distribution of errors in such cases.

Licence to exclude variables. The flip side of this rule is that a causal graph, to do the work it must do, does not need to include everything we know about a substantive domain of interest. We may know quite a lot about the causes of economic inequality, for example. But we can safely omit any other factor from the graph as long as it does not affect multiple variables in the model. Indeed, we can choose to capture any number of unspecified factors in a U term. We may be aware of a vast range of forces shaping

Figure 2.2: A DAG with unobserved confounding

whether countries democratize, but choose to bracket them for the purposes of an examination of the role of economic inequality. This bracketing is permissible as long as none of these unspecified factors also act on other variables included in the model.

You can't read functional equations from a graph. As should be clear, a DAG does not represent all features of a causal model. What it does record is which variables enter into the structural equation for every other variable: what can directly cause what. But the DAG contains no other information about the form of those causal relations. Thus, for instance, the DAG in Figure **??** tells us that M is function of both R and E, but it does not tell us whether that joint effect is additive (R and E separately increase mobilization), interactive (the effect of each depends on the value of the other), or whether either effect if linear, curvilinear or something else. This lack of information about functional forms often puzzles those encountering causal graphs for the first time; surely it would be convenient to visually differentiate, say, additive from conditioning effects. As one thinks about the variety of possible causal functions, it quickly becomes clear that there would be no simple visual way of capturing all possible functional relations. Moreover, as we shall now see, causal graphs are a tool designed with a particular analytic purpose in mind—a purpose to which we now turn.

2.2.3 Conditional independence from DAGs

If we encode our prior knowledge using the grammar of a causal graph, we can put that knowledge to work for us in powerful ways. In particular, the rules of DAG-construction allow for an easy reading of the *conditional independencies* implied by our beliefs.

To begin thinking about conditional independence, it can be helpful to conceptualize dependencies between variables as generating *flows of information*. Let us first consider a simple relationship of dependence. Returning to Figure **??**, the arrow running from I to R, implying a direct causal dependency, means that if we expect I and R to be correlated. Put differently, observing the value of one of these variables also gives us information about the value of the other. If we measured redistributive preferences, the graph implies that we would also be in a better position to infer the level of inequality, and vice versa. Likewise, I and M are also linked in a relationship of dependence: since inequality can affect mobilization (through R), knowing the the level of

inequality would allow us to improve our estimate of the level of mobilization and vice versa.

In contrast, consider I and E, which are in this graph indicated as being *independent* of one another. Learning the level of inequality, according to this graph, would give us no information whatsoever about the degree of ethnic homogeneity, and vice-versa.

Moreover, sometimes what you learn depends on *what you already know.* Suppose that we already knew the level of redistributive preferences. Would we then be in a position to learn about the level of inequality by observing the level of mobilization? According to this graph we would not: since the causal link—and, hence, flow of information between I and M—runs through R, and we already know R, there is nothing left to be learned about I by also observing M. Anything we could have learned about inequality by observing mobilization is already captured by the level of redistributive preferences, which we have already seen. In other words, if we were not to include R in the causal model, then I and M would be dependent and informative about each other. When we do include R in the causal graph, I and M are independent of one another, hence uninformative about each other. We can express this idea by saying that I and M are *conditionally independent given* R.

We say that two variables, A and C, are "conditionally independent" given a set of variables $\mathcal{B}$ if, once we have knowledge of the values in $\mathcal{B}$, knowledge of A provides no information about C and vice-versa. Taking $\mathcal{B}$ into account thus "breaks" any relationship that might exist unconditionally between A and C.

To take up another example, suppose that war is a cause of both military casualties and price inflation, as depicted in Figure **??**. Casualties and inflation will then be (unconditionally) correlated with one another because of their shared cause. If I learn that there have been military casualties, this information will lead me to think it more likely that there is also war and, in turn, price inflation (and vice versa). However, assuming that war is their only common cause, we would say that military casualties and price inflation are *conditionally independent given war.* If we already know that there is war, then we can learn nothing further about the level of casualties (price inflation) by learning about price inflation (casualties). We can think of war, when observed, as blocking the flow of information between its two

consequences; everything we would learn about inflation from casualties is already contained in the observation that there is war. Put differently, if we were just to look at cases where war is present (i.e., if we hold war constant), we should find no correlation between military casualties and price inflation; likewise, for cases in which war is absent.

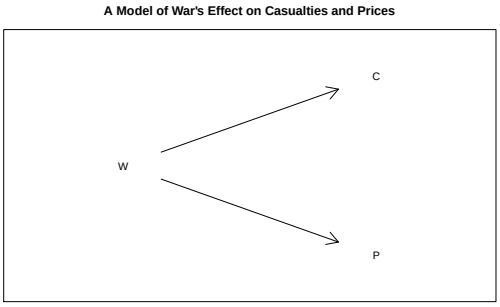

Figure 2.3: This graph represents a simple causal model in which war (W) affects both military casualties (C) and price inflation (P).

Relations of conditional independence are central to the strategy of statistical control, or covariate adjustment, in correlation-based forms of causal inference, such as regression. In a regression framework, identifying the causal effect of an explanatory variable, X, on a dependent variable, Y, requires the assumption that X's value is conditionally independent of Y's potential outcomes (over values of X) given the model's covariates. To draw a causal inference from a regression coefficient, in other words, we have to believe that including the covariates in the model "breaks" any biasing correlation between the value of the causal variable and its unit-level effect.

As we will explore, however, relations of conditional independence are of more general interest in that they tell us, given a model, *when information about one feature of the world may be informative about another feature of the world, given what we already know.* By identifying the possibilities for learning, relations of conditional independence can thus guide research design.

To see more systematically how a DAG can reveal conditional independencies, it is useful spell out three pairs of features of the flow of information in causal graphs:

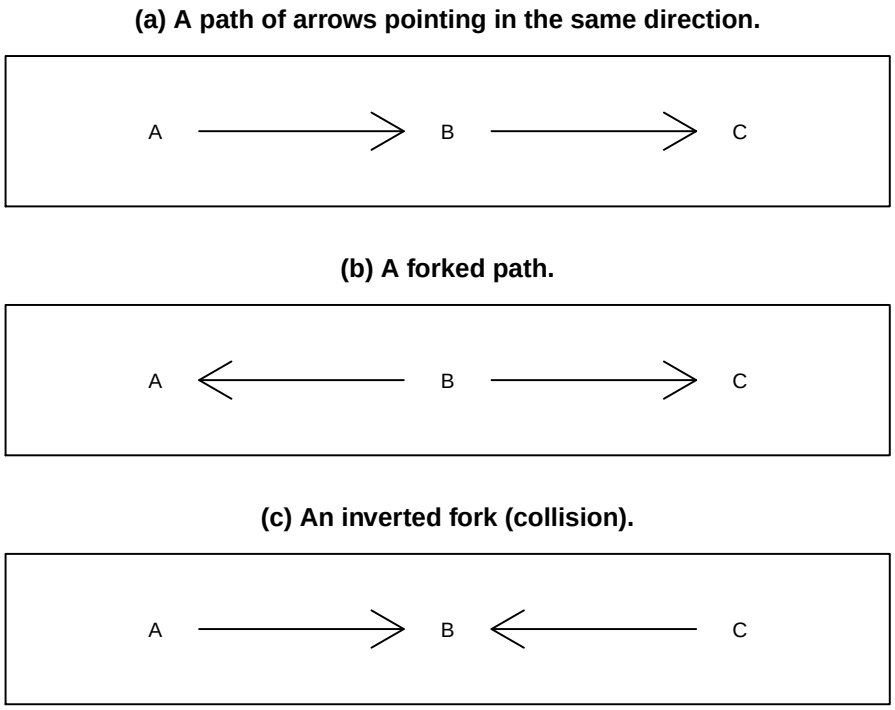

Figure 2.4: Three elementary relations of conditional independence.

(1a) Information can flow unconditionally along a path of arrows pointing in the same direction. In Panel 1 of Figure **??**, information flows across all three nodes. Learning about any one will tell us something about the other two.

(1b) Learning the value of a variable along a path of arrows pointing in the same direction *blocks* flows of information across that variable. Knowing the value of B in Panel 1 renders A no longer informative about C, and vice versa: anything that A might tell us about C is already captured by the information about B.

(2a) Information can flow unconditionally across the branches of any forked path. In Panel 2 learning only A can provide information about C and vice-versa.

(2b) Learning the value of the variable at the forking point blocks *flows* of information across the branches of a forked path. In Panel 2, learning A provides no information about C if we already know the value of B.[11]

(3a) When two or more arrowheads collide, generating an inverted fork, there is no unconditional flow of information between the incoming sequences of arrows. In Panel 3, learning only A provides no information about C, and vice-versa.

(3b) Collisions can be sites of *conditional* flows of information. In the jargon of causal graphs, B in Panel 2 is a "collider" for A and C.[12] Although information does not flow unconditionally across colliding sequences, it does flow across them *conditional* on knowing the value of the collider variable or any of its downstream consequences. In Panel 2, learning A *does* provide new information about C, and vice-versa, *if* we also know the value of B (or, in principle, the value of anything that B causes).

The last point is somewhat counter-intuitive and warrants further discussion. It is easy enough to see that, for two variables that are correlated unconditionally, that correlation can be "broken" by controlling for a third variable. In the case of collision, two variables that are *not* correlated when taken by themselves *become* correlated when we condition on (i.e., learn the value of)

[11] Readers may recognize this statement as the logic of adjusting for a confound that is a cause of both an explanatory variable and a dependent variable in order to achieve conditional independence.

[12] In the familial language of causal models, a collider is a child of two or more parents.

a third variable, the collider. The reason is in fact quite straightforward once one sees it: if an outcome is a joint function of two inputs, then if we know the outcome, information about one of the inputs can provide information about the other input. For example, if I know that you have brown eyes, then learning that your mother has blue eyes makes me more confident that your father has brown eyes.

Looking back at our democratization DAG in Figure **??**, M is a collider for R and E, its two inputs. Suppose that we again have the functional equation $M = RE$. Knowing about redistributive preferences alone provides no information whatsoever about ethnic homogeneity since the two are determined independently of one another. On the other hand, imagine that you already know that there was no mobilization. Now, if you observe that there *were* redistributive preferences, you can figure out the level of ethnic homogeneity: it must be 0. (And likewise in going from homogeneity to preferences.)

Using these basic principles, conditional independencies can be read off any DAG. We do so by checking every path connecting two variables of interest and ask whether, along those paths, the flow of information is open or blocked, given any other variables whose values are already observed. Conditional independence is established when *all* paths are blocked given what we already know; otherwise, conditional independence is absent.

2.2.4 A simple running example

We will illustrate these core ideas with a simple running example of a model of government corruption and survival.

We begin with two binary features of context. Consider, first, that a country may or may not have a free press (X). Second, the country's government may or may not be sensitive to public opinion (S).[13] Let us then stipulate what follows from these conditions. The government will engage in corruption ($C = 1$) unless it is sensitive to public opinion and there is a free press. Moreover, if and only if there is both government corruption and a free press, the press will report on the corruption ($R = 1$). Finally, the government will be removed from office ($Y = 1$) if it has acted corruptly and this gets

[13]Government sensitivity here can be thought of as government sophistication (does it take the actions of others into account when making decisions?) or as a matter of preferences (does it have a dominant strategy to engage in corruption?).

reported in the press; otherwise, the government remains in office.

As a set of equations, this simple causal model may be written as follows:

$C = 1 - X \times S$	Whether the government is corrupt
$R = C \times X$	Whether the press reports on corruption
$Y = C \times R$	Whether the government is removed from office

One thing that these equations make clear is that the variables in our model function in various places as causal-type nodes for one another. For instance, we can see from equation for C that the causal effect of a free press (X) on corruption (C) depends on whether the government is sensitive to public opinion (S): S determines C's response to X (as does X for S's effect on C). A similar relationship holds for C and X in their effect on R and for C and R in their effect on Y. As we will see below, the model also implies more complex causal-type relationships. We can, further, substitute through the causal processes to write down the functional equation for the outcome in terms of the two initial causal variables: $Y = (1 - S)X$.[14]

Let us, further, allow the two primary causal variables—the existence of a free press and the existence of a sensitive government—to vary probabilistically. In particular, we represent the probability of a free press with the population parameter λ_1^X and the probability of a sensitive government with the parameter λ_1^S.

Note that in this model, only the most "senior" specified variables, X and S, have a stochastic component (i.e., λ_1^X and λ_1^S lie between 0 and 1). All other endogenous variables are deterministic functions of other specified variables (put differently: each node has only a single nodal type).

The corresponding causal diagram for this model is shown in Figure **??**.

In later chapters we will develop this model and use it to illustrate different estimands and different strategies for case level inference.

[14]In Boolean notation (but preserving a structural equation interpretation), where Y stands for the occurrence of government removal, $Y = \neg S \wedge X$; and the function for the outcome "government retained" can be written $\neg Y = (S \wedge X) \vee (S \wedge \neg X) \vee (\neg S \wedge \neg X)$ or, equivalently, $\neg Y = S + \neg S \neg X$.

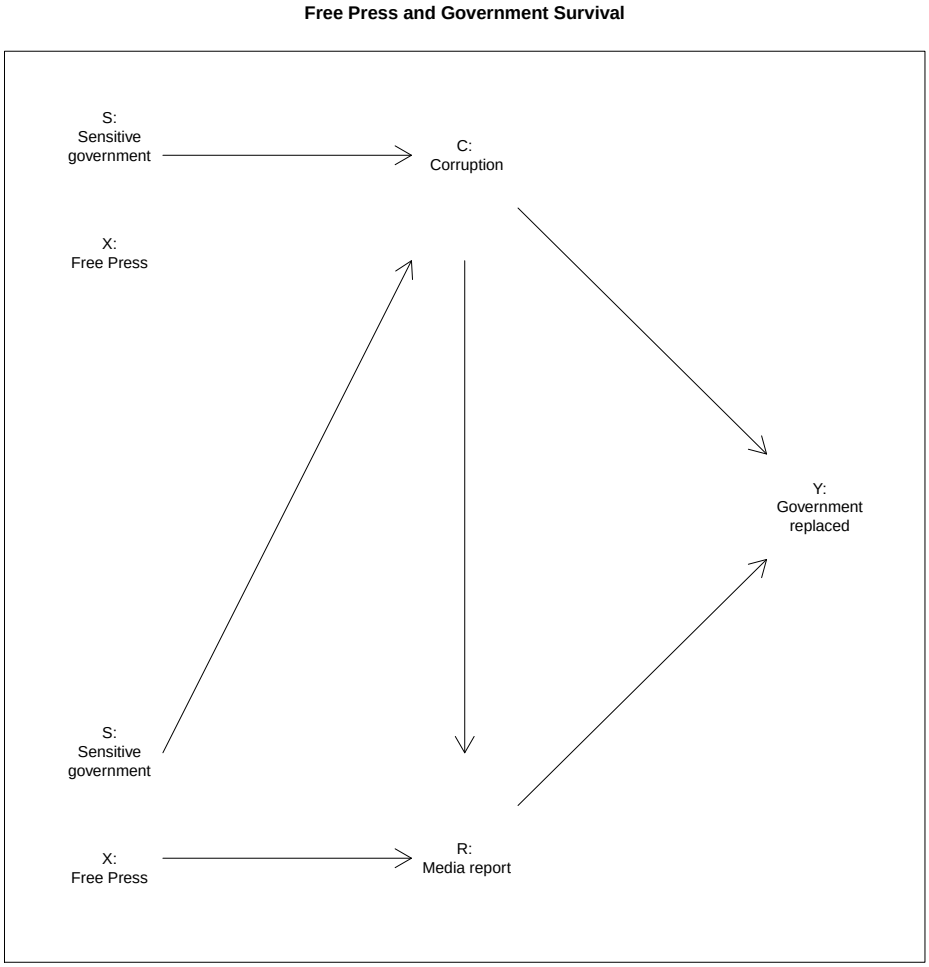

Figure 2.5: The figure shows a simple causal model. S and X are stochastic, other variables fully determined by their parents, as shown in bottom right panel.

2.3 Illustrations

We can provide more of a sense of how one might encode prior knowledge in a causal model by asking how we might construct models in light of extant scholarly works. We undertake this exercise here for three well-known works in comparative politics and international relations: Pierson's seminal book on welfare-state retrenchment (**?**); Elizabeth Saunders' research on leaders' choice of military intervention strategies (**?**); and Przeworski and Limongi's work on democratic survival (**?**), an instructive counterpoint to Boix's (**?**) argument about a related dependent variable. For each, we represent in the form of a causal model the causal knowledge that we might plausibly think we take away from the work in question. Readers might represent these knowledge bases differently; our present aim is merely to illustrate how causal models are constructed, rather than to defend a particular representation (much less the works in question) as accurate.

2.3.1 Welfare state reform: Pierson (1994)

The argument in Pierson's 1994 book *Dismantling the Welfare State?* challenged prior notions of post-1980 welfare-state retrenchment in OECD countries as a process driven primarily by socioeconomic pressures (slowed growth, rising unemployment, rising deficits, aging populations) and the rise of market-conservative ideologies (embodied, e.g., the ascendance of Thatcher and Reagan). Pierson argues that socioeconomic and ideological forces put retrenchment on the policy agenda, but do not ensure its enactment because retrenchment is a politically perilous process of imposing losses on large segments of the electorate. Governments will only impose such losses if they can do so in ways that allow them avoid blame for doing so—by, for instance, making the losses hard to perceive or responsibility for them difficult to trace. These blame-avoidance opportunities are themselves conditioned by the particular social-program structures that governments inherit.

While the argument has many more specific features (e.g., different program-structural factors that matter, various potential strategies of blame-avoidance), its essential components can be captured with a relatively simple causal model. We propose such a model in graphical form in Figure **??**. Here, the outcome of retrenchment (R) hinges on whether retrenchment

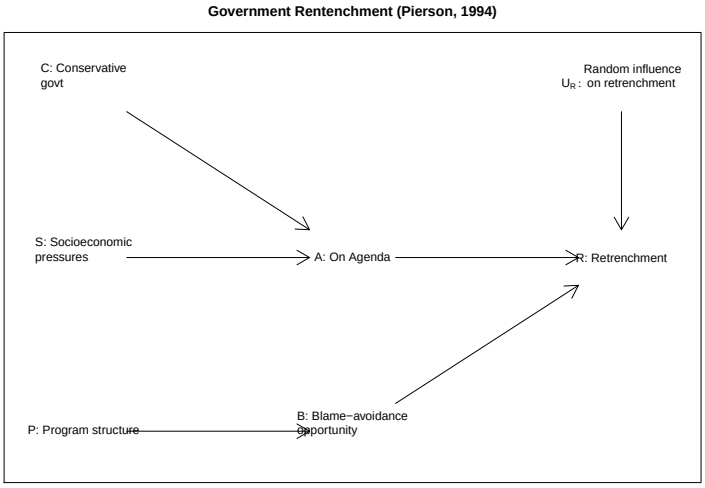

Figure 2.6: A graphical representation of Pierson (1994).

makes it onto the agenda (A) and on whether blame-avoidance strategies are available to governments (B), and on some unspecified random input (θ^R). Retrenchment emerges on the policy agenda as a consequence of both socioeconomic developments (S) and the ascendance of ideologically conservative political actors (C). Inherited program structures (P), meanwhile, determine the availability of blame-avoidance strategies.

A few features of this graph warrant attention. As we have discussed, it is the omitted arrows in any causal graph that imply the strongest statements. The graph in Panel (a) implies that C, S, P, and θ^R—which are neither connected along a directed path nor downstream from a common cause—are independent of one another. This implies, for instance, that whether conservatives govern is independent of whether program structures will allow for blame-free retrenchment. Thus, as Pierson argues, a Reagan or Thatcher can come to power but nonetheless run up against an opportunity structure that would makes retrenchment politically perilous. Further, in this graph any effect of program structures on retrenchment *must* run through their effects on blame-avoidance opportunities. One could imagine relaxing this restriction by, for instance, drawing an arrow from P to A: program structures might additionally affect retrenchment by conditioning the fiscal costliness of the welfare state, thus helping to determine whether reform makes it onto the agenda.

Where two variables *are* connected by an arrow, moreover, this does not imply that a causal effect will always operate. Consider, for instance, the arrow pointing from A to R. The fact that A sometimes affects R and sometimes does not is, in fact, central to Pierson's argument: conservatives and socioeconomic pressures forcing retrenchment on the agenda will *not* generate retrenchment if blame-avoidance opportunities are absent.

The graph also reflects a choice about where to begin. We could, of course, construct a causal account of how conservatives come to power, how socioeconomic pressures arose, or why programs were originally designed as they were. Yet it is perfectly permissible for us to bracket these antecedents and start the model with C, S, and P, as long as we do not believe that these variables have any antecedents in common. If they do have common causes, then this correlation should be captured in the DAG.[15]

The DAG itself tells us about the possible direct causal dependencies but is silent on the ranges of and functional relations among the variables. How might we express these? With three endogenous variables, we need three functions indicating how their values are determined. Moreover, every variable pointing directly into another variable must be part of that second variable's function. Let us assume all variables are binary, with each condition either absent or present. We can capture quite a lot of Pierson's theoretical logic with the following quite simple functional equations:

- $A = CS$, implying that retrenchment makes it on the agenda if and only if both conservatives are in power and socioeconomic pressures are high.
- $B = P$, implying that blame-avoidance opportunities arise when and only when program structures take a particular form
- $R = ABU_R$.

This last functional equation requires a little bit of explanation. Here we are saying that retrenchment will only occur if retrenchment is on the agenda and blame-avoidance opportunities are present (as the expression zeroes out if either of these are 0). Yet even if both are present, the effect on retrenchment also hinges on the value of θ^R. θ^R thus behaves as a causal-type variable with

[15] In DAG syntax, this correlation can be captured by placing the common cause(s) explicitly on the graph or by drawing a dashed line between the correlated nodes, leaving the source of the correlation unspecified.

respect to the effect of an AB combination on R and allows for two possible types. When $\theta^R = 1$, the AB combination has a positive causal effect on retrenchment. When $\theta^R = 0$, AB has no causal effect: retrenchment will not occur regardless of the presence of AB. A helpful way to conceptualize what θ^R is doing is that is capturing a collection of features of a case's context that might render the case susceptible or not susceptible to an AB causal effect. For instance, Pierson's analysis suggests that a polity's institutional structure might widely diffuse veto power such that stakeholders can block reform even when retrenchment is on the agenda and could be pursued without electoral losses. We could think of such a case as having a θ^R value of 0, implying that AB has no causal effect. A $\theta^R = 1$ case, with a positive effect, would be one in which the government has the institutional capacity to enact reforms that it has the political will to pursue.

2.3.2 Military Interventions: Saunders (2011)

? asks why, when intervening militarily abroad, do leaders sometimes seek to transform the *domestic* political institutions of the states they target but sometimes seek only to shape the states' external behaviors. Saunders' central explanatory variable is the nature of leaders' causal beliefs about security threats. When leaders are "internally focused," they believe that threats in the international arena derive from the internal characteristics of other states. Leaders who are "externally focused," by contrast, understand threats as emerging strictly from other states' foreign and security policies. These basic worldviews, in turn, affect the cost-benefit calculations they make about intervention strategies, via two mechanisms. Most simply, these beliefs affect perceptions of the likely security gains from a transformative intervention strategy. In addition, these beliefs affect the kinds of strategic capabilities in which leaders invest, which in turn effects the costliness and likelihood of success of alternative intervention strategies. Calculations about the relative costs and benefits of different strategies then shape the choice between a transformative and non-transformative approach to intervention. Yet leaders can, of course, only choose one of these options if they decide to intervene at all. The decision about whether to intervene depends, in turn, on at least two kinds of considerations. A leader is more likely to intervene against a given target when the nature of the dispute makes the leader's preferred strategy—given their causal beliefs—appear feasible in this situation; yet leaders may also be pushed to intervene by international or domestic audiences.

Figure **??** depicts the causal dependencies in Saunders' argument in DAG form. Working from left to right, we see that causal beliefs (C) affect the expected net relative benefits of the two strategies (B) both via a direct pathway and via an indirect pathway running through preparedness investments (P). Characteristics of a given target state or dispute (T) likewise influence B. The decision about whether to intervene (I) is then a function of three factors: causal beliefs (C), the expected relative net benefits of the strategies (B), and audience pressures (A). Finally, the choice of strategy (S) is a function of whether or not intervention occurs at all (I), cost-benefit comparisons between the two strategies (B), and other, idiosyncratic factors that may operate in various cases (θ^S).

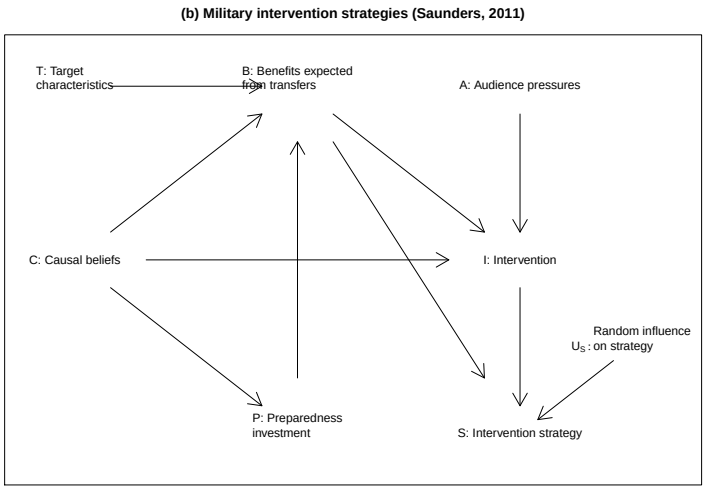

Figure 2.7: A graphical representation of Saunders' (2011) argument.

This relatively complex DAG illustrates how readily DAGs can depict the multiple pathways through which a given variable might affect another variable, as with the multiple pathways linking C to I and B (and, thus, all of its causes) to S. In fact, this graphical representation of the dependencies in some ways throws the multiplicity of pathways into even sharper relief than does a narrative exposition of the argument. For instance, Saunders draws explicit attention to how causal beliefs operate on expected net benefits via both a direct and indirect pathway, both of which are parts of an indirect pathway from C to the outcomes of interest, I and S. What is a bit easier to miss without formalization is that C also acts *directly* on the choice to

intervene as part of the feasibility logic: when leaders assess whether their generally preferred strategy would be feasible if deployed against a particular target, the generally preferred strategy is itself a product of their causal beliefs. The DAG also makes helpfully explicit that the two main outcomes of interest—the choice about whether to intervene and the choice about how—are not just shaped by some of the same causes but are themselves causally linked, with the latter depending on the former.

Omitted links are also notable. For instance, the lack of an arrow between T and A suggests that features of the target that affect feasibility have no effect on audience pressures. If instead we believed, for instance, that audiences take feasibility into account in demanding intervention, we would want to include a $T \rightarrow A$ arrow.

Turning to variable ranges and functional equations, it is not hard to see how one might readily capture Saunders' logic in a fairly straightforward set-theoretic manner. All variables except S could be treated as binary with, for instance, $C = 1$ representing internally focused causal beliefs, $P = 1$ representing preparedness investments in transformation, $B = 1$ representing expectations that transformation will be more net beneficial than non-transformation, $T = 1$ meaning that a target has characteristics that make transformation a feasible strategy, and so on. Although there are two strategies, we in fact need three values for S because it must be defined for all values of the other variables—i.e., it must take on a distinct categorical value if there is no intervention at all. We could then define functions, such as:

- $B = CPT$, implying that transformation will only be perceived to be net beneficial in a case if and only if the leader has internally focused causal beliefs, the government is prepared for a transformative strategy, and the target has characteristics that make transformation feasible
- $I = (1-|B-C|)+(1-(1-|B-C|))A$, implying that intervention can occur under (and only under) either of two alternative sets of conditions: if the generally preferred strategy and the more net-beneficial strategy in a given case are the same (i.e., such that $B - C = 0$) or, when this alignment is absent (i.e., such that $|B - C| = 0$), where audiences pressure a leader to intervene.

2.3.3 Development and Democratization: Przeworski and Limongi (1997)

? argue that democratization occurs for reasons that are, with respect to socioeconomic or macro-structural conditions, largely idiosyncratic; but once a country has democratized, a higher level of economic development makes democracy more likely to survive. Economic development thus affects whether or not a country is a democracy, but only after a democratic transition has occurred, not before. Thus, unlike in **?**, democratization in Przeworski and Limongi's argument is exogenous, rather than being determined by other variables in the model. Moreover, the dynamic component of Przeworski and Limongi's argument—the fact that both the presence of democracy and the causal effect of development on democracy depend on whether a democratic transition occurred at a previous point in time—forces us to think about how to capture over-time processes in a causal model.

We represent Przeworski and Limongi's argument in the DAG in Figure **??**. The first thing to note is that we can capture dynamics by considering democracy at different points in time as separate nodes. According to the graph, whether a country is a democracy in a given period (D_t) is a function, jointly, of whether it was a democracy in the previous period (D_{t-1}) and of the level of per capita GDP in the current period, as well as of other unspecified forces (θ^{D_t}) that lie outside the model.

Second, the arrow running from GDP_{t-1} to D_t means that GDP *may* affect democracy, not that it always does. Indeed, Przeworski and Limongi's argument is that development's effect depends on a regime's prior state: GDP matters for whether democracies continue to be democracies, but not for whether autocracies go on to become democracies. The *lack* of an arrow between D_{t-1} and GDP_{t-1}, however, implies a (possibly incorrect) belief that democracy and GDP in the last period are independent of one another.

Finally, we might consider the kind of causal function that could capture Przeworski and Limongi's causal logic. In this function, GDP should reduce the likelihood of a transition *away* from democracy but not affect the probability of a transition *to* democracy, which should be exogenously determined. One possible translation of the argument into functional terms is:

$$d_t = 1(p(1 - d_{t-1}) + d_{t-1}(1 - q(1 - gdp)) > u_{D_t})$$

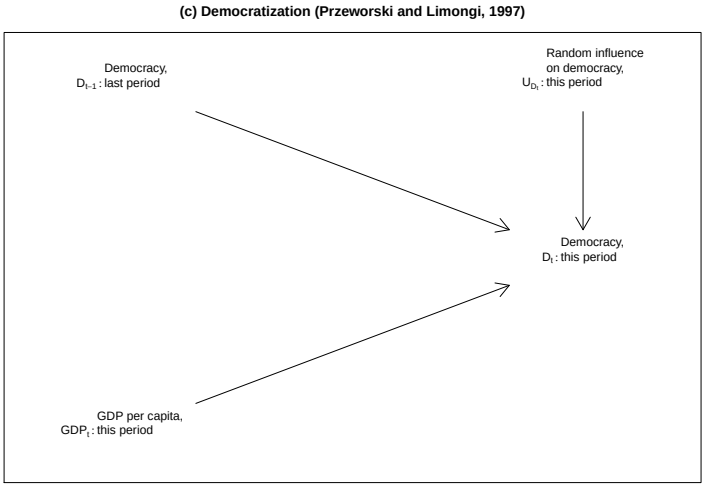

Figure 2.8: A graphical representation of Przeworski and Limongi's argument, where D_{t-1}=democracy in the previous period; GDP_t=per capita GDP in the current period; D_t=democracy in the current period.

where

- d_t and d_{t-1} are binary, representing current and last-period democracy, respectively
- p is a parameter, varying from 0 to 1, representing the probability that an autocracy democratizes
- q is a parameter, varying from 0 to 1, representing the probability that a democracy with a GDP of 0 reverts to autocracy
- gdp represents national per capita GDP, normalized on a 0 to 1 scale for the population of interest.
- θ^{D_t} represents a random, additional input into democracy with a uniform distribution on the 0 to 1 scale
- the indicator function, 1, evaluates the inequality and generates a value of 1 if and only if it is true

Unpacking the equation, the likelihood that a country is a democracy in a given period rises and falls with the expression to the left of the >-operator. This expression itself has two parts, reflecting the difference between the determinants of *transitions to* democracy (captured by the first part) and the determinants of democratic *survival* (captured by the second). The first

part comes into play—i.e., is non-zero—only for non-democracies. For non-democracies, the expression evaluates simply to p, the exogenous probability of democratization. The second part is non-zero only for democracies, where it evaluates to $1 - q$—the inverse of the reversion parameter—times $1 - gdp$: thus, the reversion probability falls as national income rises. The inequality is then evaluated by "asking" whether the expression on the left (either p or $(1 - q)gdp$) is greater than a number (θ^{D_t}) randomly drawn from a uniform distribution between 0 and 1. Thus, higher values for the expression increase the likelihood of democracy while the randomness of the θ^{D_t} threshold captures the role of other, idiosyncratic inputs.

Note how, while the functional equation nails down certain features of the process, it leaves others up for grabs. In particular, the parameters p and q are assumed to be constant for all autocracies and for all democracies, respectively, but their values are left unspecified. And one could readily write down a function that left even more openness—by, for instance, including an unknown parameter that translates GDP into a change in the probability of reversion or allowing for non-linearities, with unknown parameters, in this effect.

2.4 Chapter Appendix

2.4.1 Steps for constructing causal models

Box: **Steps for constructing causal models**

1. Identify a set of variables in a domain of interest

- You should specify the range of each variable: is it continuous or discrete?
- May include U terms representing unspecified, random influences

2. Draw a causal graph (DAG) representing beliefs about causal dependencies among these variables

- Capture direct effects only
- Arrows indicate *possible*, not constant or certain, causal effects

- The absence of an arrow between two variables indicates a belief of *no* direct causal relationship between them
- Ensure that the graph captures all correlations among variables. This means that either (a) any common cause of two or more variables is included on the graph (with implications for Step 1) or (b) correlated variables are connected with a dashed, undirected edge.

3. Write down one causal function for each endogenous variable

- Each variable's function must include all variables directly pointing into it on the graph
- Functions may take any form, as long as each set of possible causal values maps onto a single outcome value
- Functions may express arbitrary amounts of uncertainty about causal relations

4. State probabilistic beliefs about the distributions of the exogenous variables

- How common or likely to do we think different values of the exogenous variables are?
- Are they independently distributed? If in step 2 you drew an undirected edge between nodes then you believe that the connected variables are not independently distributed.

2.4.2 Model construction in code

Our `gbiqq` package provides a set of functions to implement all of these steps concisely for *binary* models – models in which all variables are dichotomous.

```
# Steps 1 and 2
# We define a model with three binary variables and specified edges between them:
model <- make_model("X -> M -> Y")

# Step 3
# Unrestricted functional forms are allowed by default, though these can
# also be reduced. Here we impose monotonicity at each step
# by removing one type for M and one for Y
model <- set_restrictions(model, labels = list(M = "10", Y="10"))
```

```
# Step 4
# We set priors over the distribution of (remaining) causal types.
# Here we set "jeffreys priors"
model <- set_priors(model, distribution = "jeffreys")

# We now have a model defined as an R object.
# Later we will ask questions of this model and update it using data.
```

These steps are enough to fully describe a binary causal model. Later in this book we will see how we can ask questions of a model like this but also how to use data to train it.

2.4.3 Test yourself! Can you read conditional independence from a graph?

As an exercise, see whether you can identify the relations of conditional independence between A and D in Figure **??**.

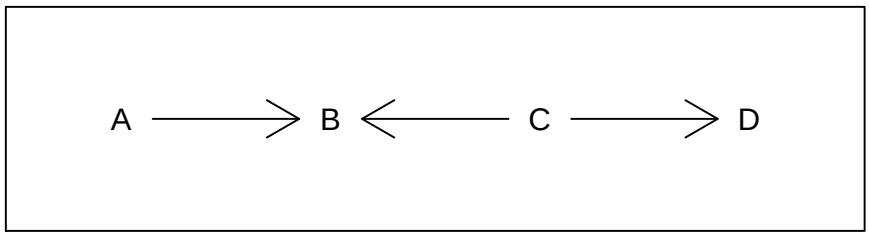

Figure 2.9: An exercise: A and D are conditionally independent, given which other variable(s)?

Are A and D independent:

- unconditionally?

Yes. B is a collider, and information does not flow across a collider if the value of the collider variable or its consequences is not known. Since no

information can flow between A and C, no information can flow between A and D simply because any such flow would have to run through C.

- if you condition on B?

No. Conditioning on a collider opens the flow of information across the incoming paths. Now, information flows between A and C. And since information flows between C and D, A and D are now also connected by an unbroken path. While A and D were independent when we conditioned on nothing, they cease to be independent when we condition on B.

- if you condition on C?

Yes. Conditioning on C, in fact, has no effect on the situation. Doing so cuts off B from D, but this is irrelevant to the A-D relationship since the flow between A and D was already blocked at B, an unobserved collider.

- if you condition on B and C?

Yes. Now we are doing two, countervailing things at once. While conditioning on B opens the path connecting A and D, conditioning on C closes it again, leaving A and D conditionally independent.

Analyzing a causal graph for relations of independence represents one payoff to formally encoding our beliefs about the world in a causal model. We are, in essence, drawing out implications of those beliefs: given what we believe about a set of direct causal relations (the arrows on the graph), what must this logically imply about other dependencies and independencies on the graph, conditional on having observed some particular set of nodes? We show in a later chapter how these implications can be deployed to guide research design, by indicating which parts of a causal system are potentially informative about other parts that may be of interest.

Chapter 3

Theories as causal models

We give an introduction to the idea of thinking of (applied) theoretical claims as claims within hierarchies of causal models. Lower level models serve as a theory for a higher level model if the higher level model can be deduced from the lower level model. The empirical content of a lower level model is the possible reduction in variance of the higher level model that it can provide.

Theory plays an important role in this book's use of causal models for causal inference. Yet the term "theory" in the empirical social sciences means very different things in different contexts. In this book, we will refer to a theory is an *explanation* of a phenomeon: a theory provides an account of how or under what conditions a set of causal relationships operate. Moreover, we can express both a theory and the claims being theorized as causal models. A theory, then, is a model that explains and implies another model—possibly with the help of some data.

We discuss toward the end of the chapter how this definition of theory relates to common understandings of theory in the social sciences. First, however, we focus on unpacking our working definition. In embedding theorization within the world of causal models, we ultimately have an empirical objective in mind. Theorizing a causal relationship of interest, in our framework, means elaborating our causal beliefs about the world in greater detail. As we show in

later chapters, theorizing in the form of a causal model allows us to generate research designs: to identify sources of inferential leverage and to explicitly and systematically link observations of components of a causal system to the causal questions we seek to answer.

3.1 Theory as a "lower-level" model

Let us say that a causal model, M', is a *theory* of M if M is implied by M'. Theory is, thus, all relative. M' might itself sit atop a theory, M'', that implies M'. To help fix the idea of theory as "supporting" or "underlying" the model(s) it theorizes, we refer to the theory, M', as a *lower*-level model relative to M and refer to M as a *higher*-level model relative to its theorization, M'.[1]

We illustrate showing two models, M', M'' that each imply a model M. In each case the lower level models contain additional nodes in a way that allows for a kind of "disaggregation" of exogenous nodes.

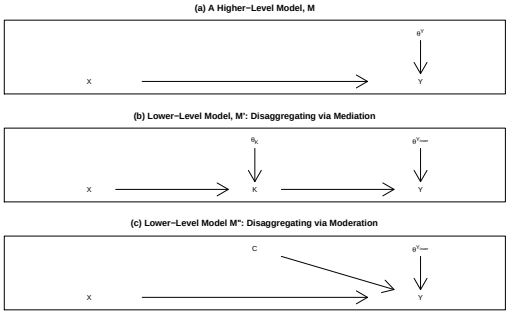

Figure 3.1: Here we represent the simple claim that one variable causes another, and two theories — lower-level models — that could explain this claim. Both model (b) and model (c) involve theorization via disaggregation of nodes.

We start with the higher-level model, M, represented in Figure **??**(a). We can then offer the model, M' in panel (b) as a *theory*, a lower-level model,

[1]We note that our definition of theory differs somewhat from that given in **?** (p207): there a theory is a (functional) causal model and a restriction over $\times_j \mathcal{R}(U_j)$, that is, over the collection of contexts envisionable. Our definition also considers probabilistic models as theories, allowing statements such as ''the average effect of X on Y is 0.5."

of M. We have added a node, K, in the causal chain between X and Y, a familiar mode of theorization. In doing do we have in fact *split* the error θ^Y into two parts: $\theta^{Y_{\text{lower}}}$ and θ^K.

Intuitively, in the higher-level model, (a), Y is a function of X and a disturbance θ^Y, the latter representing all things other than X than can affect Y. In our four-type setup, θ^Y represents all of the (unspecified) sources of variation in X's effect on Y. When we add K, X now does not directly affect Y but only does so via K. Further, we model X as acting on K "with error," with θ^K representing all of the (unspecified) factors determining X's effect on K. The key thing to notice here is that θ^K now represents *a portion of the variance that θ^Y represented in the higher-level graph*: some of the variation in X's effect on Y now arises from X's effect on K, which is captured by θ^K. So, for instance, X might have no effect on Y because θ^K takes on a value such that X has no effect on K. Likewise, any effect of X on Y must arise from an effect of X on K, captured in θ^K's value. [2] What θ^K represents, then, is that part of the original θ^Y that arose from some force other than X operating at the *first* step of the causal chain from X to Y.

So now, θ^Y is not quite the same entity in the lower-level graph that it was in the higher-level graph. In the original graph, θ^Y represented *all* sources of variation in X's effect on Y. In the lower-level model, with K as mediator, θ^Y represents only random variation in K's effect on Y. θ^Y has been expunged of any factors shaping the first stage of the causal process, which now reside in θ^K. Reflecting a convention that we use throughout the book, we highlight this change in θ^Y's meaning by referring in the second model to $\theta^{Y_{\text{lower}}}$.

Theorization here thus starts with the proliferation of substantive variables—adding beliefs about intervening steps in a causal process. But, critically, it also involves an accompanying disaggregation of unexplained variation. Addition and splitting thus go hand-in-hand: the *insertion* of a mediator between X and Y also involves the *splitting* of Y's unspecified parent (θ_Y).

Consider next model M'' panel (c) in Figure **??**, which also supports (implies) the higher-level theory in panel (a). The logical relationship between models (a) and (c), however, is somewhat different. Here the lower-level model

[2] As we emphasize further below, it is in fact only this "error" in the $X \rightarrow K$ link that makes the addition of K potentially informative as a matter of research design: if K were a deterministic function of X only, then knowledge of X would provide full knowledge of K, and nothing could be learned from observing K.

specifies one of the conditions that comprised θ^Y in the higher-level model. In specifying a moderator, C, we have extracted C from θ^Y, leaving $\theta^{Y_{\text{lower}}}$ to represent all factors *other than* C that condition X's effect on Y. (Again, the relabeling as $\theta^{Y_{\text{lower}}}$ reflects this change in the term's meaning.) While we might add a θ^C term pointing into C, this is not necessary. Whereas in Model (b) we have extracted θ^K from θ^Y, in Model (c), it is C itself that we have extracted from θ^Y, substantively specifying what had been just a random disturbance.

Critically, notice that since lower level models imply higher level models we think of theories as implying the models they are theorizing. If you believe Model M', then you also must believe Model M. If it is possible that X can affect K and possible that K can affect Y then it is possible that X can affect Y. The converse is not true, however. It is not possible to still believe that X can effect Y if you do not think that X can affect K. Similarly, if you believe Model (c), then you must also believe Model (a): if it is true that X can affect Y, possibly in ways that are moderated by C, then it is trivially true, more simply, that X can affect Y.

3.2 Illustration of unpacking causal types

We now show more specifically how causal types in lower level models map into causal types in higher level models.

For concreteness, let us return to our democratization example and consider first the very basic claim that inequality can have an affect on democratization. We represent this simple claim in Figure **??**, Panel (a). In this simple model, I may sometimes have an effect of D, and sometimes not; and that effect may be positive or negative. All of this will depend on the case's causal type.

In addition the figure shows two models that each *explain* Model (a), though in different ways. Model (b) answers the explanatory question, "*How* does inequality affect democratization?" Model (c) answers the explanatory question, "*Why* does inequality's effect on democratization vary?" Both theories provide richer, more interpretable accounts of the phenomenon of interest than the simpler model that they are theorizing.

These lower level models imply a set of causal types that are richer than that

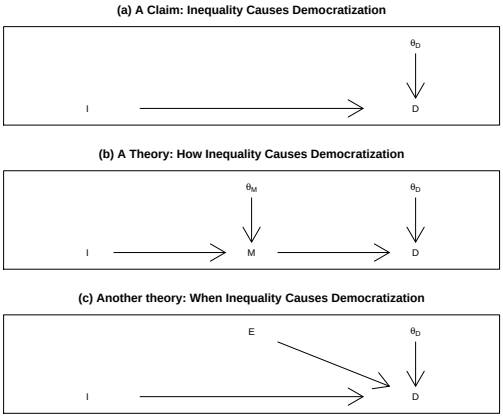

Figure 3.2: DAG representations of three theories. DAGs only capture claims that one variable causes another, conditional on other variables. Theories (b) and (c) each imply theory (a).

implied by (a). Recall that in Chapter **??**, we considered the idea that at any node, a causal type may be conceptualized as a case specific disturbance, that governs the mapping from input variables to outcome variables.

In particular if we deploy our four-causal-type function from Chapter **??** we have:

- a: $\theta^D = \theta^D_{10}$, then $D = 1 - I$ (I has a negative effect on D)
- b: $\theta^D = \theta^D_{01}$, then $D = I$ (I has a positive effect on D)
- c: $\theta^D = \theta^D_{00}$, then $D = 0$ (I has no causal effect)
- d: $\theta^D = \theta^D_{11}$, then $D = 1$ (I has no causal effect)

Knowing θ tells us how D responds to I and it ignores any heterogeneity between units as long they respond in the same way. For any causal type the model is *consistent* with I's causal effect operating for different reasons for different units, but these differences are left entirely unaccounted for.

3.2.1 Type disaggregation in a mediation model

Model (b) has causal types defined for nodes M and for D. As with the overall I, D relationship. We thus allow I to have a positive, negative, or no effect on M, with θ^M taking on four possible values, again corresponding to a, b, c, d nodal types (now: a: θ^M_{10}, b: θ^M_{01}, c: θ^M_{00}, d: θ^M_{11}).

And we allow for M to have a positive, negative, or no effect on D, with θ^D_{lower} possible values again being one of four nodal types (θ^D_{10}, θ^D_{01}, θ^D_{00}, θ^D_{11}).

We can now think about *combinations* of types in the lower-level model as mapping onto types in the higher-level model. Table **??** illustrates.

Table 3.1: Mapping from lower level nodal types on M and D to higher level causal types on D.

	$\theta^{D_{lower}}_{10}$	$\theta^{D_{lower}}_{01}$	$\theta^{D_{lower}}_{00}$	$\theta^{D_{lower}}_{11}$
θ^M_{10}	$\theta^{D_{higher}}_{01}$	$\theta^{D_{higher}}_{10}$	$\theta^{D_{higher}}_{00}$	$\theta^{D_{higher}}_{11}$
θ^M_{01}	$\theta^{D_{higher}}_{10}$	$\theta^{D_{higher}}_{01}$	$\theta^{D_{higher}}_{00}$	$\theta^{D_{higher}}_{11}$
θ^M_{00}	$\theta^{D_{higher}}_{11}$	$\theta^{D_{higher}}_{00}$	$\theta^{D_{higher}}_{00}$	$\theta^{D_{higher}}_{11}$
θ^M_{11}	$\theta^{D_{higher}}_{00}$	$\theta^{D_{higher}}_{11}$	$\theta^{D_{higher}}_{00}$	$\theta^{D_{higher}}_{11}$

For instance, in a case in which both $\theta^M = \theta^M_{01}$ (a positive effect of I on M) and $\theta^{D_{\text{lower}}} = \theta^{D_{lower}}_{01}$ (a positive effect of M on D), we have a positive effect of I on D—meaning that, in the *higher-level* model, $\theta^{D_{higher}} = \theta^{D_{higher}}_{01}$. Two linked negative effects also generate a positive effect of I on D and so map onto the same higher-level type. Further, it is easy to see that if there is no causal effect at *either* the $I \rightarrow M$ step *or* the $M \rightarrow D$ step, we will have one of the null effect types at the higher level since, in this model, I cannot affect D unless there are causal effects at both constituent steps.[3]

To foreshadow the discussion in later chapters, these mappings are critical: they allow us to use inferences drawn at a lower level to answer questions posed at a higher level.

3.2.2 Type disaggregation in a moderation model

Alternatively, we might wonder *why* inequality causes democratization. Our simple claim, in panel (a), allows that I *can* cause D, but provides no information about the conditions under which it does so. Those conditions are implicitly embedded within θ^D, where they are left unspecified. We could,

[3]These mappings, of course, hinge on the fact that I affects D *only* through M in this model (no direct effects or other pathways).

however, theorize some of what is left unsaid in in panel (a). We do this in panel (c), where we posit ethnic homogeneity (E) as a moderator of inequality's effect on democratization. Panel (c) represents a theory of panel (a) in that it can help account for variation in causal effects that is unaccounted for by the model in (a).

Model (c) has thus given substantive meaning to an aspect of the phenomenon that was merely residual variation in Model (a). Model (a) provides no account of why inequality has the effects it does, relying fully on θ^D as a placeholder for this uncertainty. In Model (c), θ^D plays a more modest role, with ethnic homogeneity doing a good deal of the work of determining inequality's possible effects.

In this graph, we again have a θ_D^{lower} term, but it is a different object from θ_D^{lower} in the mediation graph. In this moderation model, θ_D^{lower} is more complex as it determines the mapping from two binary variables into D. A causal type in this setup now represents how a case will respond to four different possible combinations of I and E values. Rather than four causal types, we now have 16, as there are 16 possible ways in which a case might respond to two binary variables (see Table **??** in Chapter **??**).

In Table **??** we give a mapping from a subset of these lower level types to the upper level types corresponding to the model in (a).

Table 3.2: Values for D given E and I. With two binary causal variables, there are 16 nodal types: 16 ways in which Y depends on I and E. These lower level types map into higher level types for a model in which Y depends on I only, as shown in the final column.

Low Type	$I=0, E=0$	$I=0, E=1$	$I=1, E=0$	$I=1, E=1$	High Type
θ^D_{0000}	0	0	0	0	θ^D_{00}
θ^D_{0001} θ^D_{0010}	0 0	0 0	0 1	1 0	θ^D_{01} if $E=1$, else θ^D_{00} θ^D_{00} if $E=1$, else θ^D_{01}
θ^D_{0011}	0	0	1	1	θ^D_{01}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
θ^D_{1110}	1	1	1	0	θ^D_{11} if $E=0$, else θ^D_{10}

Low Type	$I = 0, E = 0$	$I = 0, E = 1$	$I = 1, E = 0$	$I = 1, E = 1$	High Type
θ^D_{1111}	1	1	1	1	θ^D_{11}

Importantly we see that the mapping between lower- and higher-level types can depend on the value of the moderator. More generally, since we can think of the value of exogeneous nodes, E and I, as being nodal types for those nodes, we can think of the lower level nodal type as a concatenation of the upper level nodal types for E and D. Thus we can think of the the higher level type as depending uniquely on the fully specified lower level type.

For instance, a case that has type θ_{01} in the higher-level model if it has type θ_{0010} in the lower-level model *and* $E = 0$. This is a case for which I has a positive effect on D when $E = 0$ *and* in which E *is in fact* 0. On the other hand, the same lower level type, in combination with $E = 1$ maps onto the type θ_{10} in the higher-level model—a type in which D responds negatively to I.

In later chapters, we represent all lower- to higher-level mappings relevant to a question of interest with the use of "type-reduction" tables that allow one to readily see how inferences drawn at one level inform causal questions posed at another level.

3.3 Rules for moving between higher- and lower-level models

Thinking about models as conditionally nested within one another can be empirically useful. It provides a way of generating empirical leverage on a causal question by plumbing more deeply our background knowledge about a domain of interest. When we more fully specify higher-level claims via a more elaborate, lower-level model, we are a making explicit unspecified conditions on which the higher-level relationships depend. In doing this, we are identifying potentially observable nodes that might be informative about our research question.

As we develop lower-level models to support our claims, or determine which

claims are supported by our theories, what kinds of moves are we permitted to make? One important thing to note is that the mappings between higher-level claims and theories may not be one-to-one. A single theory can support multiple higher-level theories. Moreover, a single higher-level relation can be supported by multiple, possibly incompatible lower-level theories.

To illustrate, consider two "lower level" theories of democratization:

- (L_1): Inequality $\rightarrow$ Democratization $\leftarrow$ Mobilization
- (L_2): Inequality $\rightarrow$ Mobilization $\rightarrow$ Democratization

Note how these theories are incompatible with one another. While *Inequality* and *Democratization* are independent in L_1, they are causally related in L_2. Moreover, in L_2, *Inequality* and *Democratization* are related only through *Mobilization*, while in L_1, *Democratization* is directly affected by *Inequality*.[4]

Now, consider the following three higher-level claims:

- (H_1): Inequality $\rightarrow$ Democratization
- (H_2): Mobilization $\rightarrow$ Democratization
- (H_3): Inequality $\rightarrow$ Mobilization

H_1 could be derived from (explained by) either theory, L_1 or L_2. Although the two theories are incompatible with one another, in both theories $Inequality$ affects *Democratization*. Both theories likewise imply H_2, in which *Mobilization* affects *Democratization*.

H_3, however, can be supported only by one of these theories: only in L_2, and not in L_1, does $Inequality$ cause *Mobilization*.[5]

Thus multiple (possibly *incompatible*) theories can usually be proposed to explain any given causal effect. When seeking an explanation for, say, H_1, the choice between L_1 and L_2 is not dictated by logic; it must be drawn from

[4]Put differently, these two theories record different relations of conditional independence: in L_1, $Inequality$ and *Mobilization* are unconditionally independent, but they are not unconditionally independent in L_2. Also, in L_2, $Inequality$ is independent of *Democratization* conditional on *Mobilization*; but this is not the case in L_1.

[5]In addition, the *conditional* higher-level model $((Inequality \rightarrow Democratization)|Mobilization = 1)$ can be supported by model L_1 but not by model L_2, where holding *Mobilization* constant would sever the dependence of *Democratization* on $Inequality$.

a substantive belief about which set of causal dependencies operates in the world. On the other hand, L_2 *is* logically ruled out as an explanation of H_3. Further, any given theory logically implies multiple (necessarily *compatible*) higher-level claims about causal relations.

What, more generally, are the permissible moves across levels?

3.3.1 Moving down levels

We have already discussed two possible forms of theorization — moves down a level: (i) disaggregating existing nodes, i.e., by introducing beliefs about mediation or moderation, or (ii) adding nodes representing variation in a feature of context that is implicitly held constant in the higher-level model.

There are other possible ways of elaborating a model. For instance, we can add *antecedent conditions*: causes of nodes that were exogenous in the higher-level model. Likewise, we can add *downstream effects*: outcomes of nodes that were terminal in the higher-level model.

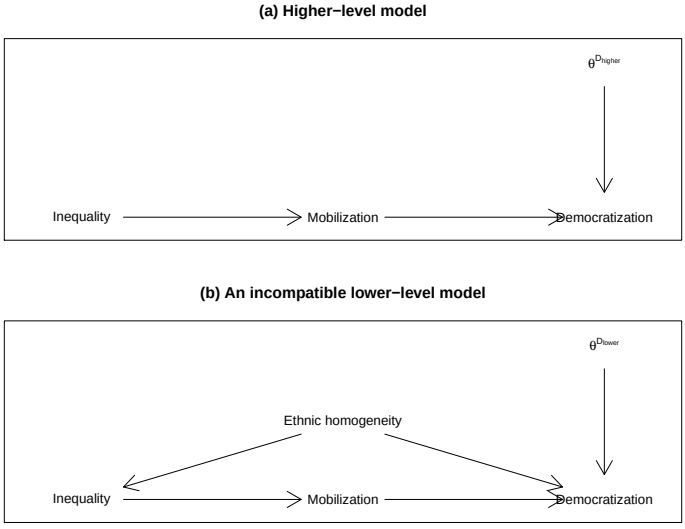

Figure 3.3: A higher-level model and a lower-level model that is impermissible.

The central principle governing allowable elaborations is that a lower-level model *must not introduce dependencies between variables that were omitted in*

the higher-level model. We provide an example of a violation of this principle in Figure **??**.

We start with a higher-level model, in panel (a), in which inequality affects democratization through mobilization. We then elaborate the model in panel (b) by adding ethnic homogeneity as a moderator of mobilization's effect. However, because ethnic homogeneity is also modeled here as affecting inequality, we have now introduced a source of dependence between inequality and democratization that was omitted from the higher-level model. In panel (a), democratization and inequality were dependent only via mobilization; and so they are conditionally independent given mobilization. In panel (b), democratization and mobilization are additionally dependent via their common cause, ethnic homogeneity. By the rules governing causal graphs (see Chapter **??**), the higher-level specifically *prohibited* this second source of dependency—since all dependencies between variables must be represented. Put differently, if two variables are independent— or conditionally independent given a third variable—in one model, then this same relation of independence (or conditional independence) must be captured in any theory of that model. A theory can *add* conditional independencies not present in the higher-level model. For instance, a mediation theory, $X \rightarrow M \rightarrow Y$, implies a conditional independence that is not present in the higher-level model that it supports, $X \rightarrow Y$: in the lower-level model only, X is conditionally independent of Y given M. But we may not theorize away (conditional) independencies insisted on by our higher-level claim.

3.3.2 Moving up levels

Moving in the other direction, what, in general, are the permissible *simplifications* of lower-level models? In other words, given a theory, what are the higher-level claims that it can support?

When we move up a level — i.e., eliminate one or more nodes — the key rule is that the higher-level graph must take into account:

(a) all *dependencies* among remaining nodes and
(b) all *variation* generated by the eliminated node.

We can work out what this means, separately, for eliminating *endogenous* nodes and for eliminating *exogenous* nodes.

Eliminating endogenous nodes

Eliminating an endogenous node means removing a node with parents (direct causes) represented on the graph. If the node also has one or more children, then the node captures a dependency: it links its parents to its children. When we eliminate this node, preserving these dependencies requires that all of the eliminated node's parents adopt—become parents of—all of the eliminated node's children. Thus, for instance in panel (b) of Figure **??**, if we were to eliminate M, M's parents (X and θ^M) need to adopt M's child, Y. We see in panel (a) of the figure, the higher-level model, that X is now pointing directly into Y.

As for θ^M, it too must now point directly into Y—though we can use a bit of shorthand to make this happen. Recall that θ^M represents the part of M that is randomly determined. Rather than drawing two separate disturbance (θ) terms pointing into Y, however, we more simply represent the combined disturbance term as θ^Y_{higher}, with the ''higher" signaling the aggregation of roots. (This is, of course, simply reversing the disaggregation that we undertook earlier to move from the higher- to the lower-level model.)

More intuitively, when we simplify away a mediator, we need to make sure that we preserve the causal relationships being mediated—both those among substantive variables and any random shocks at the mediating causal steps.[6]

Eliminating exogenous nodes

What about eliminating exogenous nodes—nodes with no parents? For the most part, exogenous nodes cannot be eliminated, but must either be replaced by or incorporated into U (or θ) terms. The reason is that we need preserve any dependencies or variation generated by the exogenous node. Figure **??** walks through four different situations in which we might want to simplify away the exogenous node, X. (Here we use the more generic U notation, though the same principles apply if these are type-receptacles(θ.)

- *Multiple children.* In (a1), we start with a lower-level model in which X has two children, thus generating a dependency between W and Y.

[6]Eliminating endogenous nodes may also operate via "encapsulated conditional probability distributions" (**?**) wherein a system of nodes, $\{Z_i\}$ is represented by a single node, Z, that takes the parents of $\{Z_i\}$ not in $\{Z_i\}$ as parents to Z and issues the children of (Z_i) that are not in (Z_i) as children. However, this is not a fundamental alteration of the graph.

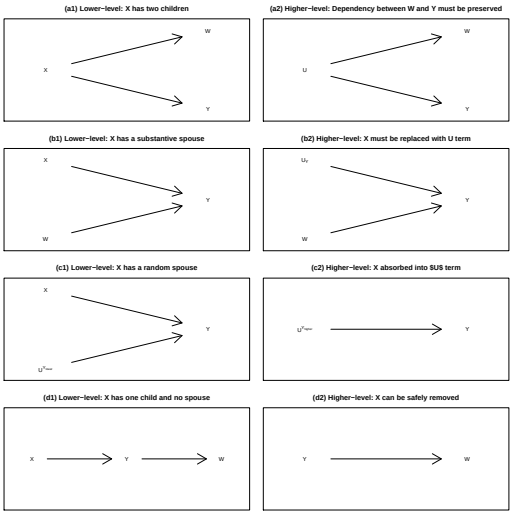

Figure 3.4: Here we represent the basic principles for eliminating exogenous nodes.

If we eliminate X, we must preserve this dependency. We can do so, as pictured in (a2), by replacing X with a U term that also points into W and Y.[7] Though we are no longer specifying what it is that connects W and Y, the correlation itself is retained.

- *Substantive spouse.* In (b1), X has a spouse that is substantively specified, W. If we eliminate X, we have to preserve the fact that Y is not fully determined by W; *something* else also generates variation in Y. We thus need to replace X with a U term, U_Y, to capture the variation in Y that is not accounted for by W.
- *U-term spouse.* In (c1), X has a spouse that is *not* substantively specified, $U^{Y_{\text{lower}}}$. Eliminating X requires, again, capturing the variance that it generates as a random input. As we already have a U term pointing only into Y, we can substitute in $U^{Y_{\text{higher}}}$, which represents both $U^{Y_{\text{lower}}}$ and the variance generated by X.[8]
- *One child, no spouse.* In (d1), X has only one child and no spouse.

[7] By DAG convention, we could, alternatively, convey the same information with a dashed, undirected line between W and Y.

[8] This aggregation cannot occur if $U^{Y_{\text{lower}}}$ also has another child, W, that is not a child of X since then we would be representing Y's and W's random components as identical, which they are not in the lower-level graph. FLAG

Here we can safely eliminate X with no loss of information. It is always understood that every exogenous node has some cause, and there is no loss of information in simply eliminating a node's causes if those causes are exogenous and do not affect other endogenous nodes in the model. In (d2) we are simply not specifying Y's cause, but we have not lost any dependencies or sources of variance that had been expressed in (d1).

One interesting effect of eliminating a substantive exogenous node can be to render seemingly deterministic relations effectively probabilistic. In moving from (b1) to (b2), we have taken a component of Y that was determined by X and converting it into a random disturbance. Just as we can explain a more probabilistic claim with a less probabilistic theory, we can derive higher-level claims with greater probabilism from theories with greater determinism.

```
## Warning in text.default(x, y + text_shift, names, cex =
## cex): font metrics unknown for character 0xa

## Warning in text.default(x, y + text_shift, names, cex =
## cex): font metrics unknown for character 0xa

## Warning in text.default(x, y + text_shift, names, cex =
## cex): font metrics unknown for character 0xa

## Warning in text.default(x, y + text_shift, names, cex =
## cex): font metrics unknown for character 0xa
```

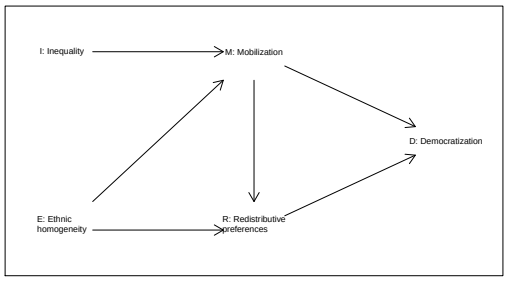

Figure 3.5: A lower-level model from which multiple higher level models can be derived.

We can apply these principles to a model of any complexity. We illustrate a

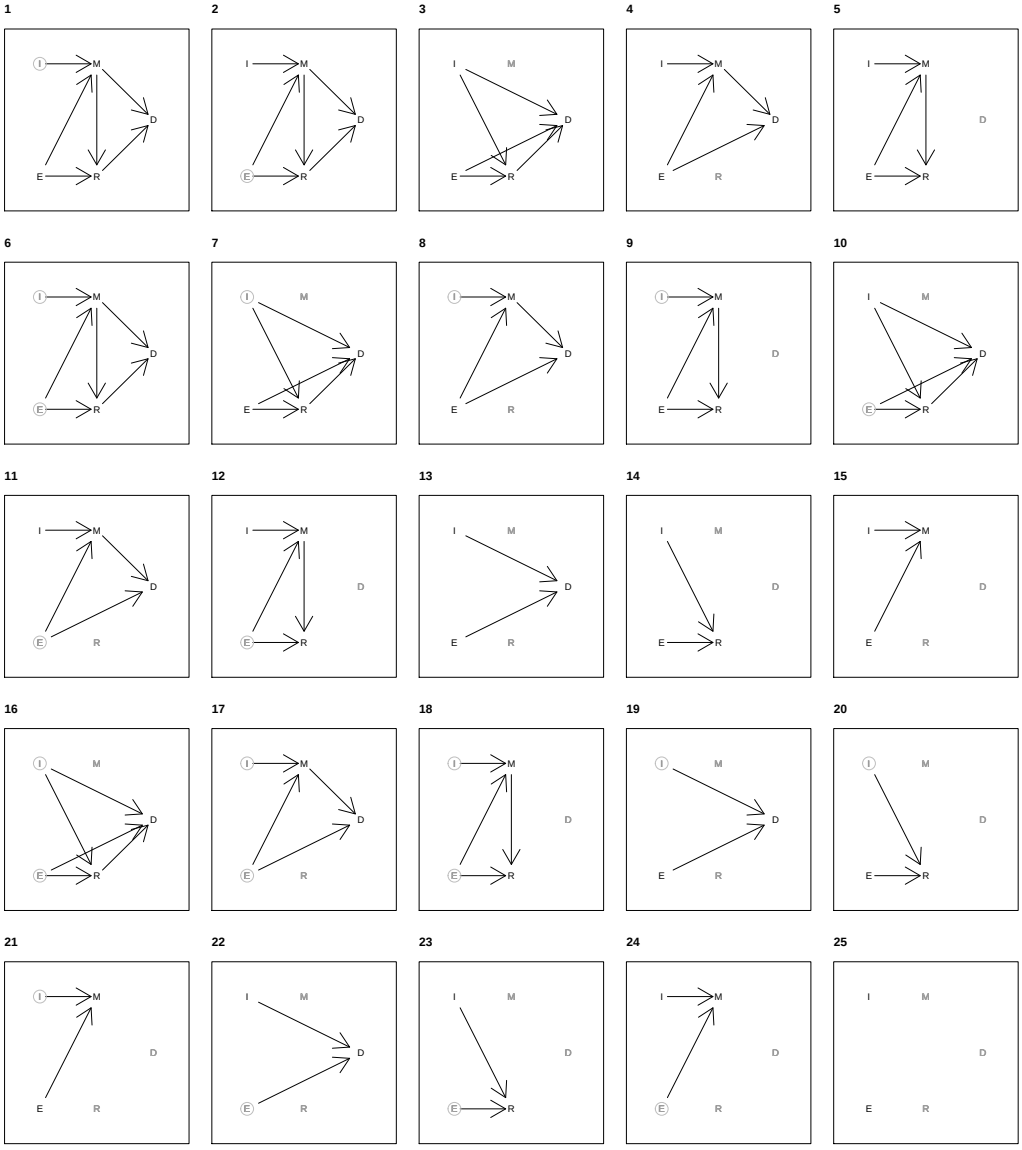

Figure 3.6: Higher level models derived from the lower level model of Figure X. Nodes that are eliminated are marked in grey; circles denote exogenous nodes that are replaced in subgraphs by unidentified variables. (A circled node pointing into two other nodes could equivalently be indicated as an undirected edge connecting the two.) Note that M, R, and D are deterministic functions of I and E in this example.

wider range of simplifications by starting with Figure **??**, which represents a somewhat amended version of our inequality and democratization model from Chapter **??**, with more complex causal relations. Then, in Figure **??**, we show all permissible reductions of the more elaborate model. We can think of these reductions as the full set of simpler claims (involving at least two nodes) that can be derived from the lower-level theory. In each subgraph,

- we mark eliminated nodes in grey;
- those nodes that are circled must be replaced with U terms; and
- arrows represent the causal dependencies that must be preserved.

Note, for instance, that neither E (because it has a spouse) nor I (because it has multiple children) can be simply eliminated; each must be replaced with a U term. Also, the higher-level graph with nodes missing can contain arrows that do not appear at all in the lower-level graph: eliminating M, for instance, forces an arrow running from X to R and another running from X to Y, as X must adopt M's children. The simplest elimination is of D itself since it does not encode any dependencies between other variables.

We can also read Figure **??** as telling us the set of claims for which the lower-level graph in Figure **??** can serve as a theory. As we can see, the range of claims that a moderately complex model can theorize is vast. For each simpler claim, moreover, there may be other possible lower-level graphs—theories besides —consistent with it.

3.3.2.1 Conditioning on nodes

A further permissible "upward" move is conditioning on a node. When we condition on a node, we are restricting the higher-level model in scope to situations in which that node's value is held constant. Doing so allows us to eliminate the node as well as all arrows pointing into it or out of it. Consider three different situations in which we might condition on a node:

- *Exogenous, with multiple children.* In simplifying (a1) in Figure **??**, we need to be sure we retain any dependence that X generates between W and Y. However, recalling the rules of conditional independence on a graph (see Chapter **??**), we know that W and Y are *independent* conditional on X. Put differently, if we restrict the analysis to contexts in which X takes on a constant value, the lower-level model implies that Y and W will be uncorrelated across cases. As fixing X's value breaks

the dependence between Y and W, we can drop X (and the arrows pointing out of it) without having to represent that dependence.

- *Exogenous, with spouse.* In simplifying (b1) or (c1) in Figure **??**, we need to account for the variation generated by X. If we fix X's value, however, then we eliminate this variation by assumption and do not need to continue to represent it (or the arrow pointing out of it) on the graph.
- *Endogenous.* When we condition on an endogenous node, we can eliminate the node as well the arrows pointing into and out of it. We, again, leverage relations of conditional independence here. If we start with graph (b) in Figure **??**, and we condition on the mediator, M, we sever the link between Y on X, rendering them conditionally independent of one another. We can thus remove M, the arrow from X to M, and the arrow from M to Y. In the new model, with M fixed, Y will be entirely determined by the random disturbance $\theta^{Y_{\text{lower}}}$.[9]

In sum, we can work with models that are simpler than our causal beliefs: we may believe a complex lower-level model to be true, but we can derive from it a sparer set of claims. There may be intervening causal steps or features of context that we believe matter, but that are not of interest for a particular line of inquiry. While these can be removed, we nonetheless have to make sure that their *implications* for the relations remaining in the model are not lost. Understanding the rules of reduction allow us to undertake an important task: checking which simpler claims are and are not consistent with our full belief set.

3.3.2.2 Relation to technical literature

Formally moving from lower level DAG to a higher level DAG requires *marginalization*: assessing the joint marginal distribution of observed nodes in the higher level graph over the distribution of unobserved nodes in the lower level graph.

Unfortunately there is no guarantee that the margin of a distribution that is consistent with a lower level DAG will be consistent with any higher level DAG (techically, "DAGS are not closed under marginalization").

[9]Note that such conditioning does not add any variance to the θ^Y term, so we retain the notation $\theta^{Y_{\text{lower}}}$.

In response, various types of richer graphs have been developed, such as "acyclic directed mixed graphs" (ADMGs) Maximal Ancestral Graphs (Spirtes and Richardson, Ancestral graph Markov models, 2002), or mDAGs (Evans 2015 (Graphs for Margins)). See also (Wermuth 2011)

ADMGs for example have directed and bidirected edges but no directed cycles and are closed under marginalization.

We use two approaches in our applications to engage with this problem. First we generally allow for *unobserved confounding* in models. Second we will allow for the estimation of lower level models with unobserved nodes.

3.4 Conclusion

We close this chapter by considering how the understanding of theory that we work with in this book compares to other prominent understandings of theory.

Theory as tautology. The claim that the number of Nash equilibria is generically odd in finite games is often understood to be a theoretical claim. Unless there are errors in the derivation of the result, the claim is true in the sense that the conclusions follow from the assumptions. There is no evidence that we could go looking for in the world to assess the claim. The same can be said of the theoretical claims of many formal models in social sciences; they are theoretical deductions of the if-then variety (**?**). Theory in this sense is true by tautology. By contrast, theory as we define it in this book refers to claims with *empirical* content: a theory refers to causal relations in the world that might or might not hold, and is susceptible to empirical testing. The deductive *logical* relations that hold in a causal model are those of conditional independence, as discussed in Chapter **??**: for instance, if X causes Y only through M in a theory, then X and Y are conditionally independent given some value of M.

Theory as a collection of maps. According to **?**, building on a semantic view of theory (**?**), a theory is a collection of models, together with a set of hypotheses linking them to the real world. As in our usage, Clarke and Primo see theories and models as very similar objects: for them, a theory is a system of models; for us, a theory is a supporting model. In both frameworks, there is no real difference in kind between models and theories.

Our approach also shares with Clarke and Primo the idea that models are not full and faithful reflections of reality; they are maps designed for a particular purpose. In the case of causal models, the purpose is to capture relationships of independence and possible causal dependence. As we have shown, that is a purpose that allows for the stripping away of detail—though it also forbids certain simplifications (such as any simplification that removes a dependency between variables). Clarke and Primo see models as useful to the extent that they are similar to features of the real world in ways related to the model's purpose. Along these lines, a causal model will be useful to the extent that it posits relations of independence that are similar to those prevailing in the domain under investigation.

Theory as a testable claim In the hypothetico-deductive framework, often traced back to **?** and highly influential in empirical political science, empirical social science is an activity of theory-*testing*. Having developed a theory, we then derive from it a set of empirical predictions and then test those predictions against evidence. In **?**, we also seek to confirm theories by developing and testing hypotheses about the similarity of a model or theory to particular features of the world. In both cases, a theory is posited—possibly on the basis of logic or background knowledge—and then assessed. The value (truth or usefulness) of the model itself is the object of inquiry.

In a causal-model framework, theories are always tentative, and we can subject any model or theory to empirical evaluation, a task to which we turn in Chapter **??**. However, in the book's setup, theories are first and foremost *expressions of what we already know and don't know* about a given causal domain when inquiry begins. We encode this background knowledge in order to inform research-design choices and draw inferences from the data. Models and theories are thus, in this sense, the world within which inquiry unfolds. Indeed, as we explore in Chapter **??**, the very questions we ask live within—can be represented as parts of—our theories.

Theory as generalization In another of the many uses of "theory," political scientists often think of theorization as generalization. For **?** and **?**, for instance, theories are by their nature general statements that we can use to explain specific events. In this view, "Diamond resources caused Sierra Leone's civil war" is a case-specific explanation; "Natural resource endowments cause civil war" is a theoretical formulation.

In our treatment of theory as a lower-level causal model, however, there is

no generic sense in which a theory is more or less general than the higher-level claim that it explains. In this book's framework, we *can* theorize by generalizing: when we elaborate a model by building in variation in a factor that was held constant in the higher-level claim, we are making the model more general in scope. If our natural resources claim implicitly applies only to weak states, we can theorize this claim by allowing state strength to vary and articulating how the natural-resource effect hinges on that claim.

However, when we theorize by disaggregating nodes—say, by adding intervening causal steps—we have in fact made a more *specific* claim. Natural resources may cause civil war under a broad set of circumstances. Natural resources will cause civil war *through looting by rebel groups* under an almost certainly narrower set of circumstances. Here, the more elaborate argument—the theorization of *why* X causes Y—is actually a stronger claim, with narrower scope, than the simpler one that it supports.

The value of parsimony ? and **?** also express a common view in characterizing *parsimony* as a quality of good theory. While they recognize that parsimony must often be traded off against other goods, such as accuracy and generality, *ceteris paribus* a more parsimonious theory—one that uses fewer causal variables to explain variation in a given outcome—is commonly understood to be a better theory.

We do not take issue with the idea that simpler models and explanations are, all else equal, better. But the succeeding chapters also demonstrate a distinctive and important way in which all else will often not be equal when we seek to use theory to guide research design and support causal inference. To foreshadow the argument to come, the elaboration of more detailed, lower-level models can direct us to new opportunities for learning. As we unpack a higher-level claim, we will often be identifying additional features of a phenomenon the observation of which can shed light on causal questions of interest. Moreover, our background beliefs—the prior knowledge on which causal inference must usually rest—are often more informative at lower levels than at higher levels: it will, for instance, often be easier for us express beliefs about causal effects for smaller steps along a causal chain than about an overarching $X \rightarrow Y$ effect.

Making things more complicated, of course, still makes things more complicated. And we should avoid doing so when the payoff is small, as it will sometimes be. But in the pages to come, we will also see a distinct set of

benefits that arise from drilling more deeply into our basis of prior knowledge when formulating inferential strategies.

3.4.1 Quantifying the gains of a theory

What are the gains of a thoery that introduces a node K relative to one that does not include K. What is the value added of the more elaborate theory?

One approach to assessing the contribution of a theory is to calculate the mean reduction in Bayes risk:

$$\text{Gains from theory} = 1 - \frac{E_{K|W}(Var(Q|K,W))}{Var(Q|W)}$$

This is a kind of R^2 measure (see also **?**).

Another approach is to ask: how much better are my guesses now compared to what I would have guessed before, given what I know now.

Expected wisdom.

$$Wisdom = \int (q_0 - q)^2 - (q_k - q)^2 p(q|k) dq$$

This captures how much better off we are with the guess we have made given current data (q_k) compared to the guess we woudl have made without it (q_0), knowing what we know now ($p(q|k)$. An advantage of this conceptualization is that you can record gains in learning even if posterior variance is larger than prior variance. Even still the implications for strategy are the same since wisdom is maximized by a strategy that reduces expected squared error.

Other possible measures of gains from theory might include the simple correlation between K and Q, or entropy-based measures (see **?** for many more possibilities).

For this problem the correlation is given by (see appendix):

$$\rho_{KQ} = \frac{(\phi_b + \phi_d)(1-2p)(p(1-p))^{.5}}{(p\phi_b + (1-p)\phi_d)(1-(p\phi_b + (1-p)\phi_d)))^{.5}}$$

One might also use a measure of "mutual information" from information theory:

$$I(Q,K) = \sum_q \sum_k P(q,k) \log \left(\frac{P(q,k)}{P(q)P(k)} \right)$$

To express this mutual information as a share of variation explained, we could divide $I(Q,K)$ by the entropy of Q, $H(Q)$ where $H(Q) = -\sum_q P(q)\log(P(q))$. The resulting ratio can be interpreted as 1 minus the ratio of the entropy of Q conditional (on K) to the unconditional entropy of Q.

For this example, Figure **??** shows gains as a function of ϕ_b given a fixed value of ϕ_d. The figure also shows other possible measures of probative value, with, in this case, the reduction in entropy tracking the reduced posterior variance closely.

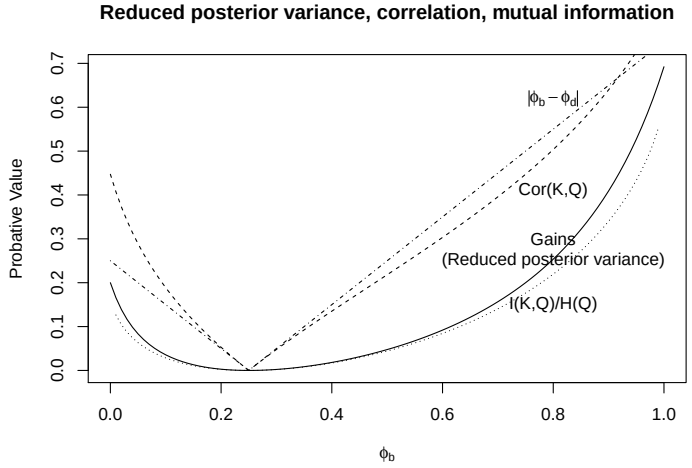

Figure 3.7: The solid line shows gains in precision (reduced posterior variance) for different values of ϕ_b given $\phi_d = 0.25$ and $p = .5$ for the example given in the text. Additional measures of probative value are also provided including $|\phi_b - \phi_d|$, the correlation of K and Q, and the reduction in entropy in Q due to mutual information in Q and K.

3.5 Chapter Appendices

3.5.1 Summary Boxes

BOX 1

Two kinds of theories.

Theories are "lower-level" causal models that explain or provide an account of a "higher-level", simpler model. There are two forms of theorization:

1. The disaggregation of nodes. A single node in a higher-level model can be split into multiple nodes. For instance, for a higher-level model in which $X \rightarrow Y \leftarrow \theta^Y$:

- *Mediation*: A mediator, M, can be introduced between X and Y, thus splitting θ^Y into θ^M and $\theta^{Y_{\text{lower}}}$. The mediation theory thus explains the $X \rightarrow Y$ relationship.
- *Moderation*: A component of θ^Y can be extracted and specified as a substantive variable. This variable is now a substantively conceptualized moderator of the $X \rightarrow Y$ relationship. The moderation theory thus provides a fuller explanation of why X has different effects on Y in different contexts.

2. Generalization. A feature of context omitted and implicitly held constant in a higher-level model can be explicitly included in the model. The higher-level model is now explained as a special case of a more general set of causal relations.

BOX 2

Rules for moving between levels

Moving down levels:

All (conditional) independencies represented in a higher-level model must be preserved in the lower-level model.

When we disaggregate or add nodes to a model, new conditional independencies can be generated. But any variables that are independent or conditionally independent (given a third variable) in the higher-level model must also be independent or conditionally independent in the lower-level model.

Moving up levels: We can move up levels by eliminating an exogenous node, eliminating an endogenous node, or conditioning on a node. When we eliminate a node from a model, we must preserve any variation and dependencies that it generates:

1. When eliminating an endogenous node, that node's parents adopt (become direct causes of) that node's children.
2. When eliminating an exogenous node, we must usually replace it with a U term. If the node has more than one child, it must be replaced with a U term pointing into both children (or an undirected edge connecting them) to preserve the dependency between its children. If the node has a spouse, the eliminated node's variation must also be preserved using a U term. Where the spouse is (already) a U term with no other children, U terms can be combined.

3. Since conditioning on a node "blocks" the path through which it connects its children, we can simply eliminate the node and the arrows between it and its children.
4. An exogenous node with no spouse and only one child can be simply eliminated.

3.5.2 Illustration of a Mapping from a Game to a DAG

Our running example supports a set of higher level models, but it can also be *implied* by a lower level models. Here we illustrate with an example in which the lower level model is a game theoretic model, together with a solution.[10]

In Figure **??** we show a game in which nature first decides on the type of the media and the politician – is it a media that values reporting on corruption

[10]Such representations have been discussed as multi agent influence diagrams, for example in **?** or **?** on "settable systems"— an extension of the "influence diagrams" described by **?**.

or not? Is the politician one who has a dominant strategy to engage in corruption or one who is sensitive to the risks of media exposure? In the example the payoffs to all players are fully specified, though for illustration we include parameter b in the voter's payoffs which captures utility gains from sacking a politician that has had a negative story written about them *whether or not they actually engaged in corruption*. A somewhat less specific, though more easily defended, theory would not specify particular numbers as in the figure, but rather assume ranges on payoffs that have the same strategic implications.

The theory is then the game plus a solution to the game. Here for a solution the theory specifies subgame perfect equilibrium.

In the subgame perfect equilibrium of the game; marked out on the game tree (for the case $b = 0$) the sensitive politicians do not engage in corruption when there is a free press – otherwise they do; a free press writes up any acts of corruption, voters throw out the politician if indeed she is corrupt and this corruption is reported by the press.

As with any structural model, the theory says what will happen but also what *would* happen if things that should not happen indeed happened.

To draw this equilibrium as a DAG we include nodes for every action taken, nodes for features that determine the game being played, and the utilities at the end of the game.

If equilibrium claims are justified by claims about the beliefs of actors then these could also appear as nodes. To be clear however these are not required to represent the game or the equilibrium, though they can capture assumed logics underlying the equilibrium choice. For instance a theorist might claim that humans are wired so that whenever they are playing a "Stag Hunt" game they play "defect." The game and this solution can be represented on a DAG without reference to the beliefs of actors about the action of other players. However, if the *justification* for the equilibrium involves optimization given the beliefs of other players, a lower level DAG could represent this by having a node for the game description that points to beliefs about the actions of others, that then points to choices. In a game with dominant strategies, in contrast, there would be no arrows from these beliefs to actions.

For our running example, nodes could usefully include the politician's expectations, since the government's actions depend on expectations of the actions

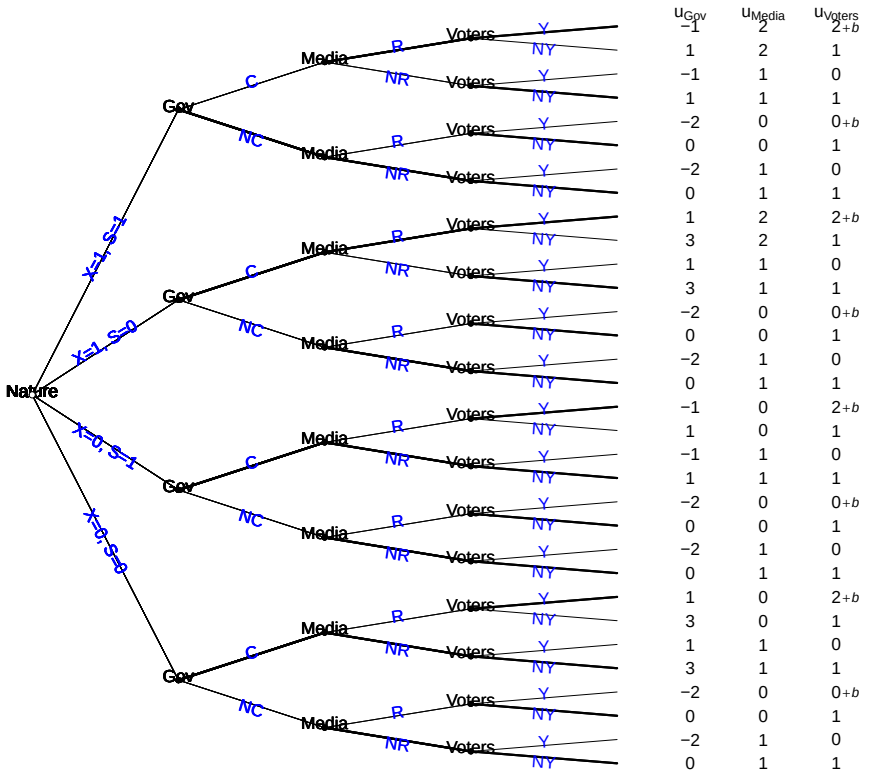

Figure 3.8: A Game Tree. Solid lines represent choices on the (unique) equilibrium path of the subgames starting after nature's move for the case in which $b = 0$.

of others. However, given the game there is no gain from including the media's expectations of the voter's actions since in this case the media's actions do not depend on expectations of the voters actions then these expectations should be included.

In Figure **??** we provide two examples of DAGs that illustrate lower level models that support our running example.

The upper panel gives a DAG reflecting equilibrium play in the game described in Figure **??**. Note that in this game there is an arrow between C and Y even though Y does not depend on C for some values of b—this is because conditional independence requires that two variables are independent for *all* values of the conditioning set. For simplicity also we mark S and X, along with b as features that affect which subgame is being played—taking the subgames starting after Nature's move. Note that the government's expectations of responses by others matters, but the expectations of other players do not matter given this game and solution. Note that the utilities appear twice in a sense. They appear in the subgame node, as they are part of the definition of the game–though here they are the utilities that players expect at each terminal node; when they appear at the end of the DAG they are the utilities that actually arise (in theory at least).

The lower level DAG is very low and much more general, representing the theory that in three player games of complete information, players engage in backwards induction and choose the actions that they expect to maximize utility given their beliefs about the actions of others. The DAG assumes that players know what game is being played ("Game"), though this could also be included for more fundamental justification of behavioral predictions. Each action is taken as a function of the beliefs about the game, the expectations about the actions of others, and knowledge of play to date. The functional equations—not shown—are given by optimization and belief formation assuming optimization by others.

These lower level graphs can themselves provide clues for assessing relations in the higher level graphs. For instance, the lower level model might specify that the value of b in the game affects the actions of the government only through their beliefs about the behavior of voters, E. These beliefs may themselves have a stochastic component, U_E. Thus b high might be thought to reduce the effect of media on corruption. For instance if $b \in \mathbb{R}_+$, we have $C = 1 - FG(1 - \mathbb{1}(b > 1))$. If X is unobserved and one is interested in

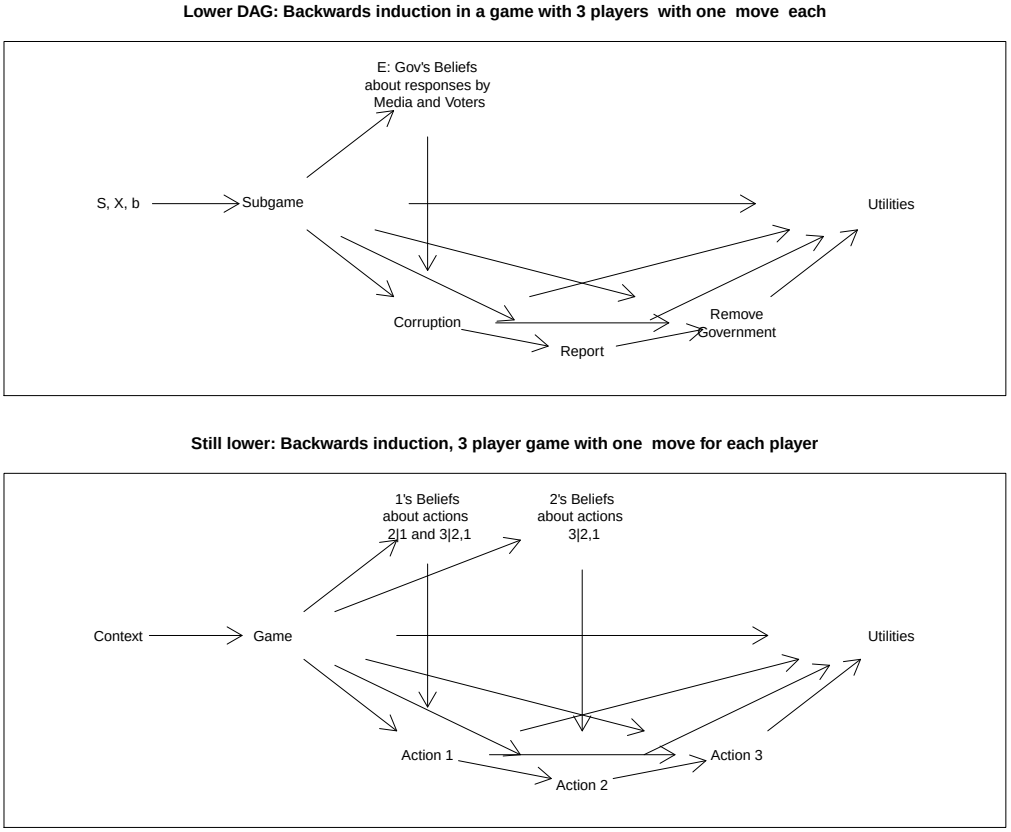

Figure 3.9: The upper panel shows a causal graph that describes relations between nodes suggested by analysis of the game in Figure **??** and which can imply the causal graph of Figure **??**. The game itself (or beliefs about the game) appear as a node, which are in turn determined by exogneous factors. The lower panel represents a still lower level and more general theory "players use backwards induction in three step games of complete information."

whether $S = 0$ caused corruption, knowledge of b is informative. It is a root node in the causal estimand. If $b > 1$ then $S = 0$ did not cause corruption. However if b matters only because of its effect on E then the query depends on U_E. In this case, while knowing b is informative about whether $S = 0$ caused $C = 1$, knowing E from the lower level graph is more informative.

Note that the model we have examined here involves no terms for U_C, U_R and U_Y—that is, shocks to outcomes given action. Yet clearly any of these could exist. One could imagine a version of this game with "trembling hands," such that errors are always made with some small probability, giving rise to a much richer set of predictions. These can be represented in the game tree as moves by nature between actions chosen and outcomes realized. Importantly in a strategic environment such noise could give rise to different types of conditional independence. For instance say that a Free Press only published its report on corruption with probability π^R, then with π^R high enough the sensitive government might decide it is worth engaging in corruption even if there is a free press; in this case the arrow from X to C would be removed. Interestingly in this case as the error rate rises, R becomes less likely, meaning that the effect of a S on Y becomes gradually weaker (since governments that are not sensitive become more likely to survive) and then drops to 0 as sensitive governments start acting just like nonsensitive governments.

Chapter 4

Causal Questions

Although a lot of empirical work focuses on identifying average causal effects, there is a rich array of other well defined causal questions that can be asked about how variables relate to each other causally. We decribe major families of question and illustrate how these can all be described as questions about the values of nodes in a causal model.

The study of causation is central to most empirical social science, whether quantitative analyses of large sets of cases or qualitative, small-N case studies. Yet a general interest in causality masks tremendous heterogeneity in the kinds of causal questions that scholars tend to ask.

Returning to our inequality and democratization example, we might seek, for instance, to know inequality's average impact on democratization across some set of cases. Alternatively, we might be interested in a particular case—say, Mongolia in 1995—and want to know whether this is a context in which inequality has an effect—a question about causal effects at the case level. Relatedly—but distinctly—we might wonder whether the level of democracy in Mongolia in 1995 is causally attributable to the level of inequality in that case. And we may be interested in *how* causal effects unfold, inquiring about the pathway or mechanism through which inequality affects democratization—a question we can also ask at two levels. We can

ask whether inequality affected democratization in Mongolia through mobilization of the masses; and we can ask how commonly inequality affects democratization through mobilization across a broad set of cases.

Rather separate methodological literatures have been devoted to the study of average causal effects, the analysis of case-level causal effects and explanations, and the identification of causal pathways. It is typically understood that their analysis requires quite distinct sets of tools. In this chapter, we take a key integrative step in showing that each of these queries can be readily captured in a causal model. More specifically, we demonstrate how causal queries can be represented as question about one or more *nodes* on a causal graph. When we assimilate our causal questions into a causal model, we are placing what we want to know in formal relation to both what we *already* know and what we can potentially *observe*. As we will see in later chapters, this move allows us then to deploy the model to generate strategies of inference: to determine which observations, if we made them, would be likely to yield the greatest leverage on our query, given our prior knowledge about the way the world works. And by the same logic, once we see the evidence, this integration allows us to "update" on our query—figure out in systematic fashion what we *have* learned—in a manner that takes background knowledge into account.

In the remainder of this chapter, we walk through the conceptualization and causal-model interpretation of five key causal queries:

- Case-level causal effects
- Case-level causal attribution
- Case-level explanation
- Average causal effects
- Causal pathways

These five are not exhaustive of the causal questions that can be captured in causal graphs, but they are among the more common foci of social scientific investigation.

4.1 Case-level causal effects

The simplest causal question is whether some causal effect operates in an individual case. Does X have an effect on Y in this case? For instance, is Yemen in 1995 a case in which a change in economic inequality would produce a change in whether or not the country democratizes? We could put the question more specifically as a query about a causal effect in a particular direction, for instance: Does inequality have a positive effect on democratization in the case of Yemen in 1995?

In counterfactual terms, a query about case-level causation is a question about what would happen if we could manipulate a variable in the case: if we could hypothetically manipulate X's value in the case, would Y's value also change? To ask whether a positive (or negative) effect operates for a case is to ask whether a particular counterfactual relation holds in that case. If we assume a binary setup for simplicity, to ask whether inequality has a positive effect on democratization is to ask: if we set I to 0 would D take on a value of 0, *and* if we set I to 1, would D take on a value of 1? (*Both* of these conditions must hold for I to have a positive effect on D.)

We can easily represent this kind of query in the context of a causal model. We show the DAG for such a model in Figure **??**. As introduced in Chapter **??**, θ^Y here represents the causal type characterizing Y's response to X and, if X and Y are binary, can take on one of four values: θ^Y_{10}, θ^Y_{01}, θ^Y_{00}, and θ^Y_{11} (which map onto our original a, b, c and d types). Importantly, given that the value of nodes (or variables) is allowed to vary across cases, this setup allows for θ_Y—the causal effect of X on Y—to vary across cases. Thus, X may have a positive effect on Y in one case (with $\theta^Y = \theta^Y_{01}$), X may have a negative ($\theta^Y = \theta^Y_{10}$) or no effect ($\theta^Y = \theta^Y_{00}$ or θ^Y_{11}) on Y in other cases.

In this model, then, the query, "What is X's causal effect in this case?" simply becomes *a question about the value of* θ_Y.

Interpreted as "what is the expected effect of X on Y?" the question becomes one of estimating $\Pr(\theta^Y = \theta^Y_{01}) - \Pr(\theta^Y = \theta^Y_{10})$.

Similarly in a mediation model of the form $X \rightarrow M \rightarrow Y$, like that discussed in Chapter 2, the question "What is the the expected effect of X on Y?" requires estimating

$$\Pr((\theta^M = \theta^M_{01} \& \theta^Y = \theta^Y_{01}) | (\theta^M = \theta^M_{10} \& \theta^Y = \theta^Y_{10})) - \Pr((\theta^M = \theta^M_{01} \& \theta^Y = \theta^Y_{10}) | (\theta^M = \theta^M_{10} \& \theta^Y = \theta^Y_{01}))$$

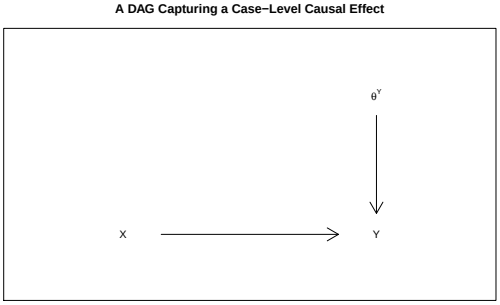

Figure 4.1: This DAG is a graphical representation of the simple causal setup in which the effect of X on Y in a given case depends on the case's causal type, represented by θ^Y. With a single binary causal variable of interest, we let θ_Y take on values θ^Y_{ij}, with i representing the value Y takes on if $X = 0$ and j representing the value Y takes on if $X = 1$. With a binary framework outcome, θ^Y ranges over the four values: θ^Y_{00}, θ^Y_{10}, θ^Y_{01} and θ^Y_{11}.

Of course, these θs are not directly observable: causal types are intrinsically unobserved properties of cases. So, as we will see in later chapters, research design becomes a challenge of determining which *observable* nodes in the graph are potentially informative about the unobservable nodes that constitute our causal queries.

4.2 Case-level causal attribution

A query about causal attribution is related to, but different from, a query about a case-level causal effect. When asking about X's case-level effect, we are asking, "*Would* a change in X cause a change in Y in this case?" The question of causal attribution is slightly different: "*Did* X cause Y to take on the value it did in this case?" More precisely, we are asking, "Given the values that X and Y *in fact* took on in this case, would Y's value have been different if X's value had been different?"

For instance, given that we know that inequality in Taiwan was relatively low and that Taiwan democratized in 1996, was low inequality a *cause* of Taiwan's democratization in 1996? Put differently, given low economic inequality and democratization in Taiwan in 1996, would the outcome in this case have been different if inequality had been high?

This goes beyond simply asking whether Taiwan is a case in which inequality has a causal effect on democratization. Whereas a case-level causal effect is defined in terms of a single θ node, we define a causal-attribution query in terms of a larger set of nodes. To attribute Y's value in a case to X, we need to know not only whether this is the kind of case in which X could have an effect on Y but also whether the context is such that X's value *in fact* made a difference.

Consider, for instance, the general setup in Figure **??**. Here, Y is a function of two variables, X and W. This means that θ^Y is somewhat more complicated than in a setup with one causal variable: θ^Y must here define Y's response to different combinations of two other variables, X and W, since *both* of these variables point directly into Y. Thus, θ^Y must cover the full set of possible causal interactions between two binary causal variables.

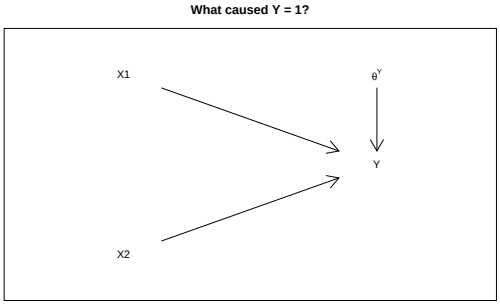

Figure 4.2: This DAG is a graphical representation of the simple causal setup in which Y depends on two variables $X1$ and $X2$. How Y responds to X1 and X2 depnds on θ^Y, the DAG itself does not provide information on whether or how X1 and X2 interact with each other.

We already saw the set of causal types for a set up like this in Chapter 2 (see Table **??**). In the table, there are four column headings representing the four possible combinations of $X1$ and $X2$ values. Each row represents one possible pattern of Y values as $X1$ and $X2$ move through their four combinations.

One way to conceptualize the size of the causal-type "space" is to note that $X1$ can have any of our four causal effects (the four binary types) on Y when $X2 = 0$; and $X1$ can have any of four causal effects when $X2 = 1$.[1] This

[1]This is precisely equivalent to noting that $X2$'s effect on Y can be of any of the four

yields 16 possible response patterns to combinations of $X1$ and $X2$ values.

A query about causal attribution—whether $X1 = 1$ caused $Y = 1$—for the model in Figure **??**, would be defined in terms of both $X2$ and θ_Y. Parallel to our Taiwan example, suppose that we have a case in which $Y = 1$ and in which $X1$ was also 1, and we want to know whether $X1$ caused Y to take on the value it did. Answering this question requires knowing whether the case's type is such that $X1$ would have had a positive causal effect on Y, *given the value of* $X2$ (which we might think of as the context). Thus, given that we start with knowledge of $X1$'s and Y's values, our query about causal attribution amounts to a query about two nodes on the graph: (a) the value of $X2$ and (b) whether the value of θ^Y is such that $X1$ has a positive causal effect given $X2$'s value.

Suppose, for instance, that we were to observe $X2 = 1$. We then need to ask whether the causal type, θ_Y, is such that $X1$ has a positive effect when $X2 = 1$. Consider type 8, or θ^{11}_{01}. This is a causal type in which $X1$ has a positive effect when $X2 = 0$ but no effect when $X2 = 1$. Put differently, $X2 = 1$ is a sufficient condition for $Y = 1$, meaning that $X1$ makes no difference to the outcome when $X2 = 1$.

In all we have four qualifying types. In other words, we can attribute a $Y = 1$ outcome to $X1 = 1$ when $X2 = 1$ and the causal type is one of these four. By parallel reasoning, there are also four causal types for which we can attribute a $Y = 1$ outcome to $X1 = 1$ when $X2 = 0$.

Thus, a question about causal attribution is a question about the *joint* value of a set of nodal types: about whether the *combination* of context and causal type is such that changing X would have changed the outcome.

4.3 Case-level explanation

So far we have been dealing with causes in the standard counterfactual sense: antecedent conditions a change in which would have produced a different outcome. Sometimes, however, we are interested in identifying antecedent conditions that were not counterfactual difference-makers but that nonetheless *generated* or *produced* the outcome. Consider, for instance, a situation

types when $X1 = 0$ and of any of the four types when $X1 = 1$.

in which an outcome was overdetermined: multiple conditions were present, each of which on their own, *could* have generated the outcome. Then none of these conditions caused the outcome in the counterfactual sense; yet one or more of them may have been distinctively important in *producing* the outcome. The concept of an *actual cause* may be useful in putting a finer point on this kind of causal question.

Let us first approach the concept at an intuitive level. An antecedent condition, A, that played a role in generating an outcome might not be a counterfactual cause because, had it not occurred, some second chain of events set in motion by B would have unfolded, generating the outcome anyway. In the standard counterfactual scenario, A is not a counterfactual cause: take away A and the outcome still happens because of the chain of events emanating from B. Yet let us imagine that the fact that A *did* occur *prevented* part of B's chain of consequences from unfolding and itself producing the outcome. Then let us imagine a tweaked counterfactual comparison in which we *fix* the observed fact that B's causal sequence did not fully unfold. We can then ask: *conditional on B's sequence not fully unfolding*, would A have been a counterfactual cause of the outcome? If so, then we say that A is an "actual cause"" of the outcome. We have, in a sense, identified A as distinctively important in the production of the outcome, even if it was not a case-level cause in the usual sense.

More formally, and using the definition provided by (**?**), building on (**?**) and others, we say that a condition (X taking on some value x) was an *actual cause* of an outcome (of Y taking on some value y), where x and y may be collections of events, if:

1. $X = x$ and $Y = y$ both happened
2. there is some set of variables, $\mathcal{W}$, such that if they were fixed at the levels that they actually took in the case, and if X were to be changed, then Y would change (where $\mathcal{W}$ can also be an empty set)
3. no strict subset of X satisfies 1 and 2 (there is no redundant part of the condition, $X = x$)

The definition thus describes a condition that *would* have been a counterfactual cause of the outcome if we were to imagine holding constant some set of events that in fact occurred (and that, in reality, might not have been constant if the actual cause had not in fact occurred).

A motivating example used in much of the literature on actual causes (e.g. **?**) imagines two characters, Sally and Billy, simultaneously throwing stones at a bottle. Both are great shots and hit whatever they aim at. Sally's stone hits first, and so the bottle breaks. However, Billy's stone *would* have hit had Sally's not hit, and would have broken the bottle. Did Sally's throw cause the bottle to break? Did Billy's?

By the usual definition of causal effects, neither Sally's nor Billy's action had a causal effect: without either throw, the bottle would still have broken. We commonly encounter similar situations in the social world. We observe, for instance, the onset of an economic crisis and the breakout of war—either of which would be sufficient to cause the government's downfall—but with the economic crisis occurring first and toppling the government before the war could do so. Yet neither economic crisis nor war made a difference to the outcome.

To return to the bottle example, while neither Sally's nor Billy's throw is a counterfactual cause, there is an important sense in which Sally's action obviously broke the bottle, and Billy's did not. This intuition is confirmed by applying the definition above. Consider first the question: Did Sally's throw break the bottle? Conditions 1 and 3 are easily satisfied, since Sally *did* throw and the bottle *did* break (Condition 1), and "Sally threw" has no strict subsets (Condition 3).

Condition 2 is met if Sally's throw made a difference, counterfactually speaking; and in determining this, we are permitted to condition on (to fix in the counterfactual comparison) any event or set of events that actually happened (or on on none at all). To see why Condition 2 is satisfied, we have to think of there being three steps in the process: Sally and Billy throw, Sally's or Billy's rock hits the bottle, and the bottle breaks. In actuality, Billy's stone did not hit the bottle. And conditioning on this actually occurring event (Billy's stone not hitting), the bottle would *not* have broken had Sally not thrown. From the perspective of counterfactual causation, it may seem odd to condition on Billy's stone not hitting the bottle when thinking about Sally not throwing the stone since Sally's throwing the stone was the very thing that prevented Billy from hitting the bottle. Yet Halpern argues that this is an acceptable thought experiment for establishing the importance of Sally's throw since conditioning is constrained to the actual facts of the case. Moreover, the same logic shows why Billy is not an actual cause. The reason

is that Billy's throw is only a cause in those conditions in which Sally did not hit the bottle. But because Sally *did* actually hit the bottle, we are not permitted to condition on Sally not hitting the bottle in determining actual causation. We thus cannot—even through conditioning on actually occurring events—construct any counterfactual comparison in which Billy's throw is a counterfactual cause of the bottle's breaking.

The striking result here is that there can be grounds to claim that a condition was the actual cause of an outcome even though, under the counterfactual definition, the effect of that condition on the outcome is 0. (At the same time, all counterfactual causes are automatically actual causes; they meet Condition 2 by conditioning on nothing at all, an empty set $\mathcal{W}$.) One immediate methodological implication follows: since actual causes need not be causes, there are risks in research designs that seek to understand causal effects by tracing back actual causes—i.e., the way things actually happened. If we traced back from the breaking of the bottle, we might be tempted to identify Sally's throw as the cause of the outcome. We would be right only in an actual-causal sense, but wrong in the standard, counterfactual causal sense. Chains of events that appear to "generate" an outcome are not always causes. [2]

As with other causal queries, the question "Was $X = x$ the actual cause of $Y = y$?" can be redefined as a question about which values for exogenous nodes produce conditions under which X could have made a difference. To see how, let us run through the Billy and Sally example again, but formally in terms of a model. Consider Figure **??**, where we represent Sally's throw (S), Billy's throw (B), Sally's rock hitting the bottle (H^S), Billy's rock hitting the bottle (H^B), and the bottle cracking (C). Each endogenous variable has a θ term associated with it, capturing its response to its parents. We capture the possible "preemption" effect with the arrow pointing from H^S to H^B, allowing that whether Sally's rock hits to affect whether Billy's rock hits.

Let us again imagine that Sally threw ($S = 1$), Billy threw ($B = 1$), and

[2]Perhaps more surprising, it is possible that the expected causal effect is negative but that X is an actual cause in expectation. For instance, say that 10% of the time Sally's shot intercepted Billy's shot but without hitting the bottle. In that case the average causal effect of Sally's throw on bottle breaking is -0.1 yet 90% of the time Sally's throw is an actual cause of bottle breaking (and 10% of the time it is an actual cause of non-breaking). For related discussions see **?**.

the bottle cracked ($C = 1$). Let us say that θ^{H^B} takes on a value such that (a) $H^B = 0$ whenever $H^S = 1$ (Sally's hit preempts Billy's) and (b) B has a positive effect on H^B when $H^S = 0$ (Billy's throw hits if Sally's doesn't). Further, assume that S has a positive effect on H^S. Let us finally posit that θ^C takes on a value such that $C = 1$ if H^B equals 1.[3] This is a set of θ values under which the query, "Does S have a causal effect on C?" must be answered in the negative. Similarly, this is a context in which $C = 1$ cannot be causally attributed to $S = 1$. If Sally had not thrown, then Sally's rock would not have hit the bottle, which means that Billy's rock would have hit, and the bottle would still have cracked—still, $C = 1$.

However, it is still possible that $S = 1$ was an actual cause of $C = 1$. To complete this query, we need to ask whether there is some node value that we can hold fixed at the value that it *actually* assumed in the case such that S would have a causal effect on the outcome. Fixing $B = 1$ (Billy throws) cannot help (since if Billy throws, Billy hits, and the bottle cracks anyway). However, under $S = 1$ and $B = 1$, given the θ values we have posited, $H^B = 0$: Billy's rock does not hit. If we hold constant that $H^B = 0$, then there is an "opportunity" for S to matter in that C is no longer forced to 1 (by Billy's rock hitting). But for S to matter under his scenario, something else has to be true: θ^C's value must allow for H^S to have a positive effect on C when $H^B = 0$.

Using our two-cause notation (with H^S on the horizontal axis, and H^B on the vertical), and given that we have already stipulated that $C = 1$ when $H^B = 1$, the one permissible value for θ^C is θ^{11}_{01}. This is causal type in which neither H^B nor H^S can be causal if both Billy and Sally throw: whenever one variable is 1, the other has no effect. But it is also a type in which each has a causal effect if the other is held at 0.

It is also the case, as we have said, that all counterfactual causes are actual causes. They are, quite simply, counterfactual causes when we hold *nothing* fixed ($\mathcal{W}$ is the empty set). Thus, in fact, any θ^S, θ^{H^S} and θ^C values in which S has a positive effect when $B = 1$ will do. This includes, for instance, a θ^C value in which Billy's hitting has no effect on the bottle (perhaps Billy doesn't throw hard enough!): e.g., θ^{01}_{01}. Here, Sally's throw is both a counterfactual cause and an actual cause of the bottle's cracking. The larger point is that

[3] That is, θ^C equals some value θ^{11}_{ij}, where H^S operates along the horizontal axis and H^B along the vertical and i and j can be any 0 or 1 values.

actual cause queries can, like all other causal queries, be defined as questions about the values of nodes in a causal model.

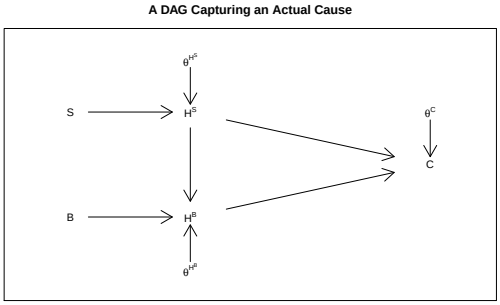

Figure 4.3: This DAG is a graphical representation of the simple causal setup in which the effect of X on Y in a given case depends on the case's causal type, represented by θ^Y. With a single binary causal variable of interest, we let θ_Y take on values θ^Y_{ij}, with i representing the value Y takes on if $X = 0$ and j representing the value Y takes on if $X = 1$. With a binary framework outcome, θ^Y ranges over the four values: θ^Y_{00}, θ^Y_{10}, θ^Y_{01} and θ^Y_{11}.

Actual causes are conceptually useful whenever there are two sufficient causes for an outcome, but one preempts the operation of the other. For instance, we might posit that both the United States' development of the atomic bomb was a sufficient condition for U.S. victory over Japan in World War II, and that U.S. conventional military superiority was also a sufficient condition and would have operated via a land invasion of Japan. Neither condition was a counterfactual cause of the outcome because both were present. However, holding constant the *absence* of a land invasion, the atomic bomb was a difference-maker, rendering it an actual cause. The concept of actual cause thus helps capture the sense in which the atomic bomb contributed to the outcome, even if it was not a counterfactual cause.

Similarly, the question of how *common* it is for a condition to be an actual cause can be expressed as values of nodes, possibly including nodes that record parameter values for the relevant exogenous nodes.

An extended notion (**?**, p 81) of actual causes restricts the imagined counterfactual deviations to states that are more likely to arise (more "normal") than the factual state. We will call this notion a ''notable cause." Similarly, one cause, A, is "more notable" than another cause, B, if a deviation in A

from its realized state is (believed to be) more likely than a deviation in B from its realized state.

For intuition, we might wonder why a Republican was elected to the presidency in a given election. In looking at some minimal winning coalition of states that voted Republican, we might distinguish between a set of states that *always* vote Republican and a set of states that usually go Democratic but voted Republican this time. If the coalition is minimal winning, then every state that voted Republican is a cause of the outcome in the standard (difference making) sense. However, only the states that usually vote Democratic are notable causes since it is only for them that the counterfactual scenario (voting Democratic) was more likely to arise than the factual scenario. In a sense, we take the "red" states' votes for the Republican as given—placing it, as it were, in the causal background—and identify as "notable" those conditions that mattered and easily could have gone differently. By the same token, we can say that, among those states that voted Republican this time, those that more commonly vote Democratic are *more* notable causes than those that less commonly vote Democratic.

Again, whether something is a notable cause, or the likelihood in some population that a condition is a notable cause, can be expressed as a claim about the value of a set of root nodes.

Though not a focus of our applied examples we show formally how to estimate these estimands in the Appendix, section XXX.

4.4 Average causal effects

A more general query asks about an average causal effect in some population. In counterfactual terms, a question about average causal effects is: if we manipulated the value of X for all cases in the population—first setting X to one value for all cases, then changing it to another value for all cases—by how much would the average value of Y in the population change? Like other causal queries, a query about an average causal effect can be conceptualized as learning about a node in a causal model.

We can do this by conceiving of any given case as being a member of a population composed of different causal types. When we seek to estimate an average causal effect, we seek information about the *shares* of these causal

types in the population.

More formally and adapted from **?**, we can use λ_{ij}^{Y} to refer to the *share* of cases in a population that has causal type θ_{ij}^{Y}. Thus, given our four causal types above, λ_{10}^{Y} is the proportion of cases in the population with negative effects; λ_{01} is the proportion of cases with positive effects; and so on. We can, of course, also think of these shares as probabilities; that is, we can think of any given case as being "drawn" from a multinomial distribution with probabilities $\lambda = (\lambda_{10}^{Y}, \lambda_{01}^{Y}, \lambda_{00}^{Y}, \lambda_{11}^{Y})$. One nice feature of this setup, with both X and Y as binary, the average causal effect can be simply characterized as the share of positive-effect cases less the share of negative-effect cases: $\lambda_{01}^{Y} - \lambda_{10}^{Y}$.

Graphically, we can represent this setup by including λ^{Y} in a more complex causal graph as in Figure **??**. As in our setup for case-level causal effects, X's effect on Y in a case depends on (and only on) the case's causal type, θ^{Y}. The key difference is that we now model the case's type not as exogenously given, but as a function of two additional variables: the distribution of causal types in a population and a random process through which the case's type is "drawn" from that distribution. We represent the type distribution as λ^{Y} (a vector of values for the proportions $\lambda_{10}^{Y}, \lambda_{01}^{Y}, \lambda_{00}^{Y}, \lambda_{11}^{Y}$) and the random process drawing a θ^{Y} value from that distribution as U_{θ}.

FLAG: CLARIFY PHILOSOPHOICAL INTERPREATAION OF LAMBDA AS SHARES

In this model, our causal query—about X's average causal effect—is thus defined by the vector λ^{Y}, and specifically by the shares of negative- and positive-causal-effect cases, respectively, in the population. What is X's average effect on Y amounts to asking: what are the values of λ_{10}^{Y} and λ_{01}^{Y}? As with θ^{Y}, λ^{Y} is not directly observable. And so the empirical challenge is to figure out what we *can* observe that would allow us to learn about λ^{Y}'s component values?[4]

We can, of course, likewise pose queries about other population-level causal quantities. For instance, we could ask for what proportion of cases in the population X has a positive effect: this would be equivalent to asking the

[4]Note also that λ^{Y} can be thought of as itself drawn from a distribution, such as a Dirichlet. The hyperparameters of this underlying distribution of λ would then represent our uncertainty over λ and hence over average causal effects in the population.

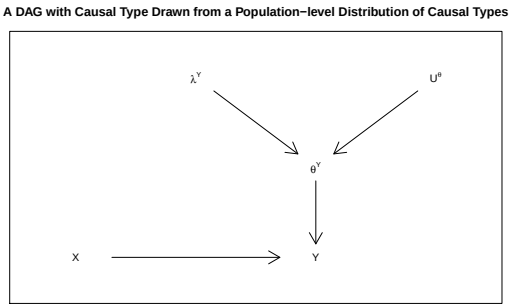

Figure 4.4: This DAG is a graphical representation of a causal setup in which cases are drawn from a population composed of different causal types. As before, X's effect on Y is a function of a causal-type variable, θ^Y. Yet here we explicitly model the process through which the case's type is drawn from a distribution of types in a population. The variable λ is a vector representing the multinomial distribution of causal types in the population while U_θ is a random variable representing the draw of each case from the distribution defined by λ. A case's causal type, θ^Y, is thus a joint function of λ^Y and U^{θ_Y}.

value of λ^Y_{01}, one element of the λ^Y vector. Or we could ask about the proportion of cases in which X has no effect, which would be asking about $\lambda^Y_{00} + \lambda^Y_{11}$.

4.5 Causal Paths

To develop richer causal understandings, researchers often seek to describe the causal path or paths through which effects propagate. Consider the DAG in Figure **??**, in which X can affect Y through two possible pathways: directly and via M. Assume again that all variables are binary, taking on values of 0 or 1. As we have seen in Chapter **??**, mediation models require causal-type nodes that point into any mediators as well as into the outcome variable. So here we have drawn in a causal-type variable defining M's response to X, θ^M, and a causal-type variable capturing Y's response, θ^Y. Importantly, θ^Y defines Y's response to *two* parent variables: M and X.

Suppose that we observe $X = 1$ and $Y = 1$ in a case. Suppose, further, that we have reasonable confidence that X has had a positive effect on Y in

this case. We may nonetheless be interested in knowing whether that causal effect ran *through* M. We will refer to this as a query about a causal path. A causal path query, of course, goes beyond assessing whether some mediating event along the path occurred. We cannot, for instance, establish that the top path in Figure **??** was operative simply by determining the value of M in this case—though that will likely be useful information.

Rather, the question of whether the top (mediated) causal path is operative is a composite question of two parts: First, does X have an effect on M in this case? Second, does that effect—the difference in M's value caused by a change in X—in turn *cause* a change in Y's value? In other words, what we want to know is whether the effect of X on Y depends on—*will not operate without*—the effect of X on M.[5] Framing the query in this way makes clear that asking whether a causal effect operated via a given path is in fact asking about a specific set of causal effects lying along that path.

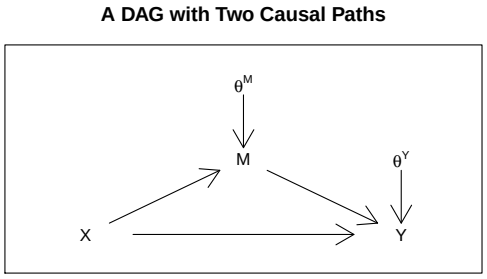

Figure 4.5: Here X has effects on Y both indirectly through M and directly.

As we can show, we can also define a causal-path query as a question about specific nodes on a causal graph. In particular, just as we have defined other questions about causal effects in terms of causal-type nodes, a causal path can also be defined in terms of the values of type nodes: specifically, in the

[5] A very similar question is taken up in work on mediation where the focus goes to understanding quantities such as the "indirect effect"" of X on Y via M. Formally, the indirect effect would be

$$Y(X = 1, M = M(X = 1, \theta^M), \theta^Y) - Y(X = 1, M = M(X = 0, \theta^M), \theta^Y))$$

, which captures the difference to Y if M were to change in the way that it would change due to a change in X, but without an actual change in X (**?**, p 132, **?**).

present example, in terms of the nodes θ^M and $theta^Y$. To see why, let us first note that there are two combinations of effects that would allow X's positive effect on Y to operate via M: (1) X has a positive effect on M, which in turn has a positive effect on Y; or (2) X has a negative effect on M, which has a negative effect on Y.

Thus, in establishing whether X affects Y through M, the first question is whether X affects M in this case. Whether or not it does is a question about the value of the causal-type node, θ^M. Let us assume that θ^M can take on four possible values corresponding to the four possible responses to X: $\theta^M_{10}, \theta^M_{01}, \theta^M_{00}, \theta^M_{11}$.[6] For sequence (1) to operate, θ^M must take on the value θ^M_{01}, representing a positive effect of X on M. For sequence (2) to operate, θ^M must take on the value θ^M_{10}, representing a negative effect of X on M.

θ^Y, as for our causal-attribution example, defines Y's response to different combinations of two other variables—here, X and M—since *both* of these variables point directly into Y. Another way to think about this setup is that M is not just a possible mediator of X's indirect effect; M is also a potential *moderator* of X's direct effect. Where X can have both an mediated effect through M and a direct effect, X and M also potentially *interact* in affecting Y.

This results in sixteeen possible values for θ^Y—again as shown above in Table **??**.

What values of θ^Y then are compatible with the operation of the M causal path? Let us first consider this question with respect to sequence (1), in which X has a positive effect on M, and that positive effect is necessary for X's positive effect on Y to occur. For this sequence to operate, θ^M must take on the value of θ^M_{01}. When it comes to θ^Y, then, what we need to look for types in which X's effect on Y *depends on M's taking on the value it does as a result of X's positive effect on M.*

We are thus looking for causal types that represent two kinds of counterfactual causal relations operating on nodes. First, X must have a positive effect on Y when M changes as it should given X's positive effect on M. Second, that change in M, generated by a change in X, must be *necessary* for X's positive effect on Y to operate. The thought experiment here thus imagines

[6] In other words, X's effect on M could be negative, positive, absent with M stuck at 0, or absent with M stuck at 1, respectively.

a situation in which X changes from 0 to 1,[7] but M does *not* change to the value that it should as a result of this change in X. We then inspect our types to see if Y would change from 0 to 1 in this situation. It is this counterfactual that isolates the causal significance of the path that runs through M. It is only if Y would *not* change to 1 in this situation that we have identified a causal-type for which the M-mediated path matters.

Assuming a positive effect of X on M ($\theta^M = \theta^M_{01}$), we thus need to apply three queries to θ^Y:[8]

1. Is $X = 1$ a counterfactual cause of $Y = 1$? Establishing this positive effect of X involves two queries:

   a) Where $X = 0$, does $Y = 0$? As we are assuming X has a positive effect on M, if $X = 0$ then $M = 0$ as well. We thus look down the $X = 0, M = 0$ column and eliminate those types in which we do not observe $Y = 0$. This eliminates types 9 through 16.

   b) Where $X = 1$, does $Y = 1$? Again, given X's assumed positive effect on M, $M = 1$ under this condition. Looking down the $X = 1, M = 1$ column, we eliminate those types where we do not see $Y = 1$. We retain only types $2, 4, 6$, and 8.

2. Is X's effect on M necessary for X's positive effect on Y? That is,

[7]This is the natural thought experiment when explaining a case with realized value of $X = 1$, in which the outcome can be thought of as having been generated by a change from $X = 0$. The identification of types does hinge, however, on the direction in which we imagine types changing. In other situations, we might observe $X = Y = 0$ and thus conceive of the outcome as having been generated by a change from $X = 1$ to $X = 0$ (again, assuming a positive effect of X on Y). When we do this, query 2 below changes: we are now looking for types in which $Y = 1$ when $X = 0$ but $M = 1$. (Does Y stay at 1 when X moves to 0 but M doesn't?) The queries are then satisfied by types 6 and 8, rather than 2 and 6.

[8]Using standard potential outcomes notation, we can express the overall query, conditioning on a positive effect of X on M, via the inequality $Y(1, M(1)) - Y(0, M(0)) > Y(1, M(0)) - Y(0, M(0))$. The three specific queries formulated below simply correspond to the three unique elements of this expression. We can also readily map the path query that we are defining here—does the positive effect of X on Y depend on X's effect on M—onto a query posed in terms of indirect effects. For instance, in our binary setup, conditioning our path query on a positive causal effect of X on Y, a positive effect of X on M, and an imagined change from $X = 0$ to $X = 1$ generates precisely the same result (identifies the same θ^Y types) as asking which θ^Y types are consistent with a positive indirect effect of X on Y, conditioning on a positive total effect and $X = 1$.

do we see $Y = 1$ *only* if M takes on the value that $X = 1$ generates $(M = 1)$? To determine this, we inspect the *counterfactual* condition in which $X = 1$ yet $M = 0$, and we ask: does $Y = 0$? Of the four remaining types, only 2 and 6 pass this test.

Under these and only these two values of θ^Y—θ^{01}_{00} and θ^{11}_{00}—we will see a positive effect of X on Y for which the M-mediated path is causally necessary as long as X also has a positive effect on M. These two θ^Y values are also different from one another in an interesting way. For type θ^{11}_{00}, X's effect on Y runs strictly through M: if M were to change from 0 to 1 *without* X changing, Y would still change from 0 to 1. X is causally important for Y *only* insofar as it affects M. In a case of type θ^{11}_{00}, then, anything else that similarly affects M would generate the same effect on Y as X does. In type θ^{01}_{00}, however, both X's change to 1 *and* the resulting change in M are necessary to generate Y's change to 1; X's causal effect thus requires both the mediated and the unmediated pathway. Andhere X itself matters in the counterfactual sense; for a case of type θ^{01}_{00}, some other cause of M would *not* generate the same effect on Y.

We can undertake the same exercise for sequence (2), in which X first has a negative effect on M, or $\theta^M = \theta^M_{10}$. Here we adjust the three queries for θ^Y to take account of this negative effect. Thus, we adjust query 1a so that we are looking for $Y = 0$ when $X = 0$ and $M = 1$. In query 1b, we look for $Y = 1$ when $X = 1$ and $M = 0$. And for query 2, we want types in which Y fails to shift to 1 when X shifts to 1 but M stays at 1. Types θ^{00}_{01} and θ^{00}_{11} pass these three tests.

In sum, we can define a query about causal paths as a query about the value of θ terms on the causal graph. For the graph in Figure **??**, asking whether X's effect runs via the M-mediated path is asking whether one of four combinations of θ^M and θ^Y hold in case:

- $\theta^M = \theta^M_{01}$ and $(\theta^Y = \theta^{01}_{00}$ or $\theta^{11}_{00})$
- $\theta^M = \theta^M_{01}$ and $(\theta^Y = \theta^{00}_{01}$ or $\theta^{00}_{11})$

It is worth noting how different this formulation of the task of identifying causal pathways is from widespread understandings of process tracing. Scholars commonly characterize process tracing as a method in which we determine whether a mechanism was operating by establishing whether the events lying along that path occurred. As a causal-model framework makes clear, finding

out that $M = 1$ (or $M = 0$, for that matter) does not establish what was going on causally. Observing this intervening step does not by itself tell us what value M *would* have taken on if X had taken on a different value, or whether this would have changed Y's value. We need instead to conceive of the problem of identifying pathways as one of figuring out the *counterfactual* response patterns of the variables along the causal chain. As we will demonstrate later in the book, explicitly characterizing those response patterns as nodes in a causal model helps us think systematically about empirical strategies for drawing the relevant inferences.

Chapter 5

Bayesian Answers

We run through the logic of Bayesian updating and show how it is used for answering causal queries. We illustrate with applications to correlational and process tracing inferences.

Bayesian methods are just sets of procedures to figure out how to update beliefs in light of new information.

We begin with a prior belief about the probability that a hypothesis is true. New data then allow us to form a posterior belief about the probability of the hypothesis. Bayesian inference takes into account the consistency of the evidence with a hypothesis, the uniqueness of the evidence to that hypothesis, and background knowledge about the problem.

In the next section we review the basic idea of Bayesian updating. The following section applies it to the problem of updating on causal estimands given a causal model and data.

5.1 Bayes Basics

For simple problems, Bayesian inference accords well with our intuitions. Once problems get slightly more complex however, our intuitions often fail

us.

5.1.1 Simple instances

Say I draw a card from a deck. The chances it is a Jack of Spades is just 1 in 52. If I tell you that the card is indeed a spade and asked you now what are the chances it is a Jack of Spades, you should guess 1 in 13. If I told you it was a heart you should guess there is no chance it is a Jack of Spades. If I said it was a face card and a spade you should say 1 in 3.

All those answers are applications of Bayes' rule. In each case the answer is derived by assessing what is possible, given the new information, and then assessing how likely the outcome of interest among the states that are possible. In all the cases you calculate:

$$\text{Probability Jack of Spades} \mid \text{Information} = \frac{\text{Is Jack of Spades Consistent with Information?}}{\text{How many cards are consistent with Information?}}$$

The same logic goes through when things are not quite so black and white.

Now consider two slightly trickier examples.

Interpreting Your Test Results. Say that you take a test to see whether you suffer from a disease that affects 1 in 100 people. The test is good in the sense that if you have the disease it will say you have it with a 99% probability. If you do not have it, then with a 99% probability, it will say that you do not have it. The test result says that you have the disease. What are the chances you have it? You might think the answer is 99%, but that would be to mix up the probability of the result given the disease with the probability of the disease given the result. In fact the right answer is 50%, which you can think of as the share of people that have the disease among all those that test positive. For example if there were 10,000 people, then 100 would have the disease and 99 of these would test positive. But 9,900 would not have the disease and 99 of these would test positive. So the people with the disease that test positive are half of the total number testing positive.

As an equation this might be written:

$$\text{Probability You have the Disease} \mid \text{Test} = \frac{\text{How many people have the disease and test positive?}}{\text{How many people test positive?}}$$

Two-Child Problem Consider last an old puzzle found described **?**. *Mr Smith has two children, A and B. At least one of them is a boy. What are the chances they are both boys?* To be explicit about the puzzle, we will assume that the information that one child is a boy is given as a truthful answer to the question "is at least one of the children a boy?" Assuming that there is a 50% probability that a given child is a boy, people often assume the answer is 50%. But surprisingly the answer is 1 in 3. The information provided rules out the possibility that both children are girls and so the right answer is found by readjusting the probability that two children are boys based on this information. As an equation:

$$\text{Probability both are boys} \mid \text{Not both girls} = \frac{\text{Probability both boys}}{\text{Probability they are not both girls}} = \frac{\text{1 in 4}}{\text{3 in 4}}$$

5.1.2 Bayes' Rule for Discrete Hypotheses

Formally, all of these equations are applications of Bayes' rule which is a simple and powerful formula for deriving updated beliefs from new data.

The formula is given as:

$$\begin{aligned} \Pr(H|\mathcal{D}) &= \frac{\Pr(\mathcal{D}|H)\Pr(H)}{\Pr(\mathcal{D})} && (5.1) \\ &= \frac{\Pr(\mathcal{D}|H)\Pr(H)}{\sum_{H'}\Pr(\mathcal{D}|H')\Pr(H'))} && (5.2) \end{aligned}$$

where H represents a hypothesis and $\mathcal{D}$ represents a particular realization of new data (e.g., a particular piece of evidence that we might observe).

Looking at the formula we see that the posterior belief derives from three considerations. First, the likelihood: how likely are we to have observed these data if the hypothesis were true, $\Pr(\mathcal{D}|H)$? Second, how likely were we to have observed these data regardless of whether the hypothesis is true

or false, $\Pr(\mathcal{D})$? These first two questions, then, capture how consistent the data are with our hypothesis and how specific the data are to our hypothesis. As shown in the equation above the second question can usefully be reposed as one about all the different ways (alternative Hypotheses, H') that could give rise to the data.

Note, that contrary to some claims, the denominator does not require a listing of all possible hypotheses, just an exhaustive collection of hypotheses. For example we might have the notion of the probability that the accused's fingerprints would be on the door if she were or were guilty without having to decompose the "not guilty" into a set of hypotheses regarding who else might be guilty.

Our posterior belief is further conditioned by the strength of our prior level of confidence in the hypothesis, $\Pr(H)$. The greater the prior likelihood that our hypothesis is true, the greater the chance that new data consistent with the hypothesis has *in fact* been generated by a state of the world implied by the hypothesis.

5.1.3 The Dirichlet family and Bayes' Rule for Continuous Parameters

This basic formula extends in a simple way to collections of continuous variables. For example, say we are interested in the value of some parameter vector θ (as a vector, θ can contain many quantities we are uncertain about), we can calculate this, given a prior probability distribution over possible values of θ, p, and given data D as:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_{\theta'} p(\mathcal{D}|\theta')p(\theta')d\theta}$$

Bayes rule requires the ability to express a prior distribution but it does not require that the prior have any particular properties other than being probability distributions.

In practice however when we are dealing with continuous parameters, it can be useful to make use of "off the shelf" distributions.

In practice we will often be interested in forming beliefs about the share of

units that are of a particular type. For this type of question we will make quite heavy use of "Dirichlet" distributions – a family of distributions that capture beliefs about shares.

Consider for example the share of people in a population that voted—this is a quantity between 0 and 1. Two people might may both believe that the turnout was around 50% but may differ in how certain they are about this claim. One might claim to have no information and to believe that any turnout rate between 0 and 100% is equally likely, giving an expected turnout of 50%; another might be completely confident that the number if 50% and entertain no other possibilities.

We can capture such beliefs quite well by using the Beta distribution—a special case of the Dirichlet. The Beta is a distribution over the $[0, 1]$ that is governed by two parameters , α and β. In the case in which both α and β are 1, the distribution is uniform – all values are seen as equally likely. As α rises large outcomes are seen as more likely and as β rises, lower outcomes are seen as more likely. If both rise proportionately the expected outcome does not change but the distribution becomes tighter.

An attractive feature of the Beta distribution is that if one has a prior Beta(α, β) over the probability of some event (e.g. that a coin comes up heads), and then one observes a positive case, the Bayesian posterior distribution is also a Beta with with parameters $\alpha + 1, \beta$. Thus in a sense if people start with uniform priors and build up knowledge on seeing outcomes, their posterior beliefs should be Beta distributions.

Figure **??** shows a set of such distributions, starting with one that has greater variance than uniform (this corresponds to the non informative "Jeffrey's prior"), then uniform, then for a case in which multiple negative and positive outcomes are seen, in equal number, and finally a set of priors with mean of 3/4.

Dirichlet distributions generalize the Beta to the situation in which there are beliefs not just over a proportion, or a probability, but over collections of probabilities. For example if four outcomes are possible and each is likely to occur with probability θ_k, $k = 1, 2, 3, 4$ then beliefs about these probabilities are distributions over the a three dimensional unit simplex—that is, all 4 element vectors of probabilities that add up to 1. The distribution has as many parameters as there are outcomes and these are traditionally recorded

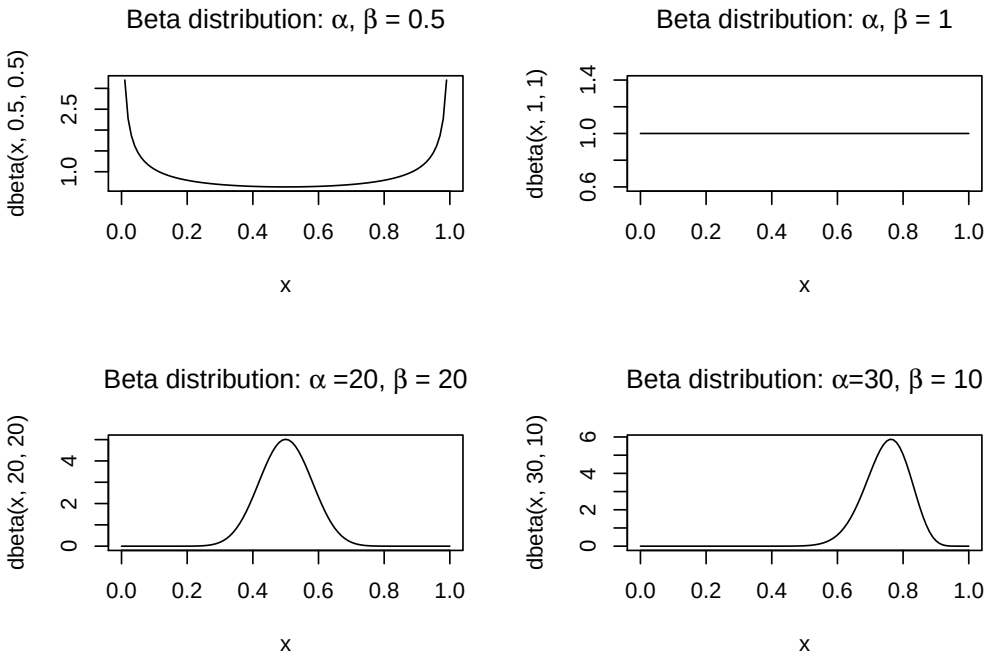

Figure 5.1: Beta distributions

in a vector, α. Similar to the Beta distribution, an uninformative prior (Jeffrey's prior) has α parameters of $(.5, .5, .5, ...)$ and a uniform ("flat") distribution has $\alpha = (1, 1, 1, , ...)$.

As with the Beta distribution, the Dirichlet updates in a simple way. If you have a Dirichlet prior with parameter $\alpha = (\alpha_1, \alpha_2, ...)$ and you observe outcome 1, for example, then then posterior distribution is also Dirichlet with parameter vector $\alpha' = (\alpha_1 + 1, \alpha_2, ...)$.

5.1.4 Moments

In what follows we often refer to the "posterior mean" or the "posterior variance." These are simply summary statistics of the posterior distribution and can be calculated easily once the posterior is known. For example the posterior mean of a parameter θ_1—just one in a collection of parameters stored in θ—is simply $\overline{\theta}_1|\mathcal{D} = \int \theta_1 p(\theta|\mathcal{D})d\theta$. Note importantly that this is calculated using the posterior over the entire vector θ, there is no notion of updating parameter θ_1 on its own. Similarly the posterior variance is $\int(\theta_1 - (\overline{\theta}_1|\mathcal{D})^2 p(\theta|\mathcal{D})d\theta$.

5.1.5 Bayes estimation in practice

Although the principle of Bayesian inference is quite simple, in practice calculating posteriors for continuous parameters is computationally complex.

In principle with continuous parameters there is an infinity of possible parameter values. Analytic solutions are not, in general, easy to come by and so in practice researchers use some form of sampling.

Imagine for instance you were interested in forming a posterior on the share intending to vote democrat, given polling data. (This is not truly continuous, but with large elections it might as well be).

One approach is to coarsen the parameter space—we calculate the probability of observing the polling data given possible values $\theta = 0, \theta = .1, \theta = .2, ..., \theta = 1$, and, apply Bayes rule to form a posterior for each of these these possibilities. The downside of the this approach is that it for a decent level of precision it becomes computationally expensive with large parameter spaces and parameter spaces get large quickly. For instance if you are interested in vote shares you might find .4, .5, and .6 too coarse and want posteriors for

0.51 or even 0.505; this would require calculations for 200 parameter values. If you had two parameters that you wanted to slice up each into 200 possible values, you would then have 40,000 parameter pairs to worry about. What's more, *most* of those calculations would not be very informative if the real uncertainty all lies in some small (though possibly unknown) range – such as between 40% and 60%.

An alternative approach is to use variants of Markov Chain Monte Carlo sampling. Under these approaches parameter vectors are sampled and their likelihood is evaluated. If they have high likelihood then new parameter vectors near them are draw with a high probability. Based on the likelihood associated with these new draws, new draws are made. The result is a chain of draws that build up to approximate the posterior distribution. The output from these procedures is not a set of probabilities for each possible parameter vector but rather a a set of draws of parameter vectors from the posterior distribution.

Many algorithms have been developed to achieve these tasks efficiently; in all of our applications we rely on the `stan` procedures which involve….

5.2 Bayes applied

5.2.1 Bayesian Inference on Queries

In Chapter 2 we described estimands of interest as queries over the values of root nodes in directed acyclic graphs.

Once queries are defined in terms of the values of roots then formation of beliefs, given data W, about estimands follows immediately from application of Bayes rule.

Let $Q(u)$ define the value of the query in context u. The updated beliefs about the query are given by the distribution:

$$P(q|W) = \int_{u:Q(u)=q} P(u|W)du = \int_{u:Q(u)=q} \frac{P(W|u)P(u)}{\int_{u'} P(W|u')P(u')du'}du$$

This expression gathers together all the contexts that produce a given value

of Q and assesses how likely these are, collectively, given the data.[1] For an abstract representation of the relations between assumptions, queries, data, and conclusions, see Figure 1 in **?**.

Return now to Mr Smith's puzzle. The two "roots" are the sexes of the two children, child A and child B. The query here is Q: "Are both boys?" which can be written in terms of the roots. The statement "$Q = 1$" is equivalent to the statement (A is a boy & B is a boy). Thus it takes the value $q = 1$ in just one context. Statement $q = 0$ is the statement ("A is a boy & B is a girl" or "A is a girl & B is a boy" or "A is a girl & B is a girl"). Thus $q = 0$ in three contexts. If we assume that each of the two children is equally likely to be a boy or a girl with independent probabilities, then each of the four contexts is equally likely. The result can then be figured out as $P(Q = 1) = \frac{1\times\frac{1}{4}}{1\times\frac{1}{4}+1\times\frac{1}{4}+1\times\frac{1}{4}+0\times\frac{1}{4}} = \frac{1}{3}$. This answer requires summing over only one context. $P(Q = 0)$ is of course the complement of this, but using the Bayes formula one can see that it can be found by summing over the posterior probability of three contexts in which the statement $Q = 0$ is true.

We will often want to think about our causal queries being collections of states of the world — i.e., of unit causal types. Returning to our discussion of queries in Chapter **??**, suppose we start with the model $X \rightarrow M \rightarrow Y$, and our query is whether X has a positive effect on Y. This is a query that is satisfied by four sets of unit types: those in which X has a positive effect on M and M has a positive effect on Y, with X being either 0 or 1; and those in which X has a negative effect on M and M has a negative effect on Y, with X being either 0 or 1. Our inferences on the query will thus involve gathering these different unit types, and their associated posterior probabilities, together.

One interesting feature of Bayesian updating is that we update more strongly in favor of the hypothesis for which the evidence is least damaging to the most-likely ways in which the hypothesis could be true. Suppose our prior belief was that it was much more unlikely that M had a negative effect on Y, than that M had a positive effect on Y. This makes one of the ways in which X could have a positive effect on Y (the chain of negative effects) much less likely than the other way in which X could have a positive effect on Y (the chain of positive effects). This means that evidence, say, against

[1] Learning about roots from observed data is sometimes termed *abduction*; see **?**, p 206.

a chain of negative effects and evidence against a chain of positive effects will not be equally consequential for our query: in particular, we will update more strongly against the query if we find evidence against a chain of positive effects than if we find evidence against a chain of negative effects. Evidence against a chain of positive effects speaks against the *most* likely way in which the query could be true, whereas evidence against a chain of negative effects speaks against a way the query could be true that we did not think was very likely to begin with.

5.2.2 Simple Bayesian Process Tracing

Process tracing in its most basic form seeks to use within case evidence to draw inferences about the case. For example, with a focus on whether X caused Y , data on a "clue", K, is used to make inference about whether or not the outcome in that case was generated by the case's treatment status. We refer to the within-case evidence gathered during process tracing as *clues* in order to underline their probabilistic relationship to the causal relationship of interest. Readers familiar with the framework in **?** can usefully think of our "clues" as akin to causal process observations, although we highlight that there is no requirement that the clues be generated by the causal process.

To make inferences, the analyst looks for clues that will be observed with some probability if the case is of a given type and that will *not* be observed with some probability if the case is *not* of that type.

It is relatively straightforward to express the logic of process tracing in Bayesian terms, a step that will aid the integration of qualitative with quantitative causal inferences. As noted by others (e.g. **?**, **?**, **?**), there is an evident connection between the use of evidence in process tracing and Bayesian inference. .

To illustrate, suppose we are interested in regime collapse. We already have X, Y data on one authoritarian regime: we know that it suffered economic crisis ($X = 1$) and collapsed ($Y = 1$). We want to know what caused the collapse. To make progress we will try to draw inferences given a "clue." Beliefs about the probabilities of observing clues for cases with different causal effects derive from theories of, or evidence about, the causal process connecting X and Y. Suppose we theorize that the mechanism through which economic crisis generates collapse runs via diminished regime capacity to reward its

supporters during an economic downturn. A possible clue to the operation of a causal effect, then, might be the observation of diminishing rents flowing to regime supporters shortly after the crisis. If we believe the theory, then this is a clue that we might believe to be highly probable for cases of type b that have experienced economic crisis (where the crisis in fact caused the collapse) but of low probability for cases of type d that have experienced crisis (where the collapse occurred for other reasons).

To make use of Bayes rule we need to:

1. define our parameters, which are the key quantities of interest
2. provide prior beliefs about the parameters of interest
3. define a likelihood function
4. provide the probability of the data
5. plug these into Bayes' rule to calculate a posterior on the parameters of interest

We discuss each of these in turn.

Parameters. The inferential challenge is to determine whether the regime collapsed *because* of the crisis (b type) or whether it would have collapsed even without it (d type). We do so using further information from the case—one or more clues. We use the variable K to register the outcome of the search for a clue, with K=1 indicating that a specific clue is searched for and found, and K=0 indicating that the clue is searched for and not found.

Let $j \in \{a, b, c, d\}$ refer to the type of an individual case. Our hypothesis, in this initial setup, consists simply of a belief about j for the case under examination: specifically whether the case is a b type ($j = b$). The parameter of interest is the causal type.

Prior. We then assign a prior degree of confidence to the hypothesis ($p = Pr(H)$). This is, here, our prior belief that an authoritarian regime that has experienced economic crisis is a b.

Likelihood. The likelihood, $\Pr(K = 1|H)$ is the probability of observing the clue, when we look for it in our case, if the hypothesis is true—i.e., here, if the case is a b type. The key feature of a clue is that the probability of observing the clue is believed to depend on the case's causal type. In order to calculate the probability of the data we will in fact need two such probabilities: we let ϕ_b denote the probability of observing the clue for a

case of b type ($\Pr(K = 1|j = b)$), and ϕ_d the probability of observing the clue for a case of d type ($\Pr(K = 1|j = d)$). The key idea in many accounts of process tracing is that the *differences* between these probabilities provides clues with ''probative value," that is, the ability to generate learning about causal types. The likelihood, $\Pr(K = 1|H)$, is simply ϕ_b.

Probability of the data. This is the probability of observing the clue when we look for it in a case, *regardless* of its type, ($\Pr(K = 1)$). More specifically, it is the probability of the clue in a treated case with a positive outcome. As such a case can only be a b or a d type, this probability can be calculated simply from ϕ_b and ϕ_d, together with our beliefs about how likely an $X = 1, Y = 1$ case is to be a b or a d type. This probability aligns (inversely) with Van Evera's concept of ''uniqueness."

Inference. We can now apply Bayes' rule to describe the learning that results from process tracing. If we observe the clue when we look for it in the case, then our *posterior* belief in the hypothesis that the case is of type b is:

$$\Pr(j = b|K = 1, X = Y = 1) = \frac{\phi_b p}{\phi_b p + \phi_d(1 - p)}$$

In this exposition we did not make use of a causal model in a meaningful way—we simply need the priors and the clue probabilities.

In fact, however, these numbers can be derived from a causal model. To illustrate, imagine a simple causal model in which the X, Y relationship is completely mediated by K. In particular, suppose, from background knowledge of the conditional distribution of outcomes given their causes, we have that:

- $\Pr(K = 1|X = 0) = 0$, $\Pr(K = 1|X = 1) = .5$
- $\Pr(Y = 1|K = 0) = .5$, $\Pr(Y = 1|K = 1) = 1$

This data is consistent with a world in which half b and c types in the first step and half b and d types in the second step. Assume that the case at hand is sampled from this world.

Then we can calculate that the prior probability, p, that X caused Y given

$X = Y = 1$ is $p = \frac{1}{3}$.[2] We can also calculate the probability that $K = 1$ for a treated b and d case respectively as $\phi_b = 1$ and $\phi_d = 0.5$ (convince yourself of these numbers!). We then get:

$$\Pr(j = b|K = 1, X = Y = 1) \quad = \quad \frac{1 \times \frac{1}{3}}{1 \times \frac{1}{3} + 0.5 \times \frac{2}{3}} = 0.5$$

We thus shift our beliefs from a prior of $\frac{1}{3}$ to a posterior of $\frac{1}{2}$. In contrast had we *not* observed the clue our posterior would have been 0.

As should be clear from the above, the inferential leverage in process tracing comes from differences in the probability of observing $K = 1$ for different causal types. Thus, the logic described here generalizes Van Evera's familiar typology of tests by conceiving of the certainty and uniqueness of clues as lying along a continuum.

Van Evera's four tests ("smoking gun," "hoop," "straw in the wind," and "doubly decisive") represent, in this sense, special cases—particular regions that lie on the boundaries of a "probative-value space." To illustrate the idea, we represent the range of combinations of possible probabilities for ϕ_b and ϕ_d as a square in Figure **??** and mark the spaces inhabited by Van Evera's tests. As can be seen, the type of test involved depends on both the relative *and* absolute magnitudes of ϕ_b and ϕ_d. The probative value of a test depends on the difference between them. Thus, a clue acts as a smoking gun for proposition "b" (the proposition that the case is a b type) if it is highly unlikely to be observed if proposition b is false, and more likely to be observed if the proposition is true (bottom left, above diagonal). A clue acts as a "hoop" test if it is highly likely to be found if b is true, even if it still quite likely to be found if it is false. Doubly decisive tests arise when a clue is very likely if b and very unlikely if not. It is, however, also easy to imagine clues with probative qualities lying in the large space amidst these extremes.[3]

[2]Given $X = 1$, $Y = 1$ is consistent with b types at both stages, which arises with probability .25, or with a d type in the second stage, which arises with probability .5. The conditional probability is therefore $.25/.75 = 1/3$.

[3]We thank Tasha Faifield for discussions around this graph which differs from that in **?** by placing tests more consistently on common rays originaity from (0,0) and (1,1).

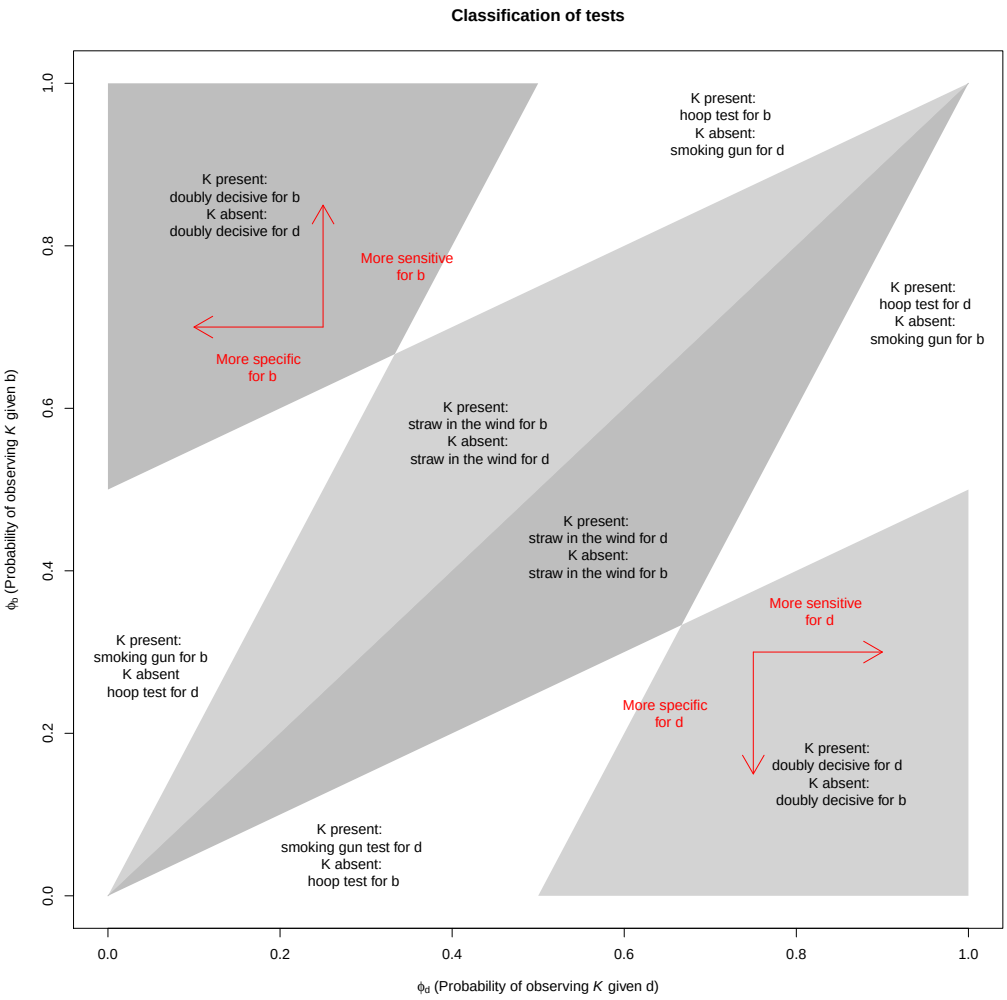

Figure 5.2: A mapping from the probability of observing a clue if the proposition that a case is a b type is true (ϕ_b) or false (ϕ_d) to a generalization of the tests described in Van-Evera (1997).

In this illustration we note that we draw both the priors and the probative value from a causal model. If we altered the model—for example if we had a stronger first stage and so a larger value for $\Pr(K = 1|X = 0)$—this would alter both our prior, p, and our calculations of ϕ_d. An implication of this is that, although one might be tempted to think of the priors and the probative values as independent quantities, and contemplate how inferences change as priors change (as we did for example in teh treatment in **?**), keeping probative value fixed, that kind of thought experiment may assume values that are not justified by an underlying model.

5.3 Three principles of Bayesian updating

FLAG: REDO THESE THREE EXAMPLES WITHOUT PHI

5.3.1 Priors matter

The amount of learning that results from a given piece of new data depends strongly on prior beliefs. We saw this already with the example of interpreting our test results above. Figure **??** illustrates the point for process tracing inferences.

In each subgraph of Figure **??** , we show how much learning occurs under different scenarios. The horizontal axis indicates the level of prior confidence in the hypothesis and the curve indicates the posterior belief that arises if we do (or do not) observe the clue. As can be seen, the amount of learning that occurs—the shift in beliefs from prior to posterior—depends a good deal on what prior we start out with. For a smoking gun test, the amount of learning is highest for values roughly in the 0.2 to 0.4 range—and then declines as we have more and more prior confidence in our hypothesis. For a hoop test, the amount of learning when the clue is *not* observed is greatest for hypotheses in which we have middling-high confidence (around 0.6 to 0.8), and minimal for hypotheses in which we have a very high or a very low level of confidence.

The implication here is that our inferences with respect to a hypothesis must be based not just on the search for a clue predicted by the hypothesis but also on the *plausibility* of the hypothesis, based on other things we know. Suppose, for instance, that we fail to observe evidence that we are 90 percent sure we *should* observe if a hypothesized causal effect has occurred: a strong

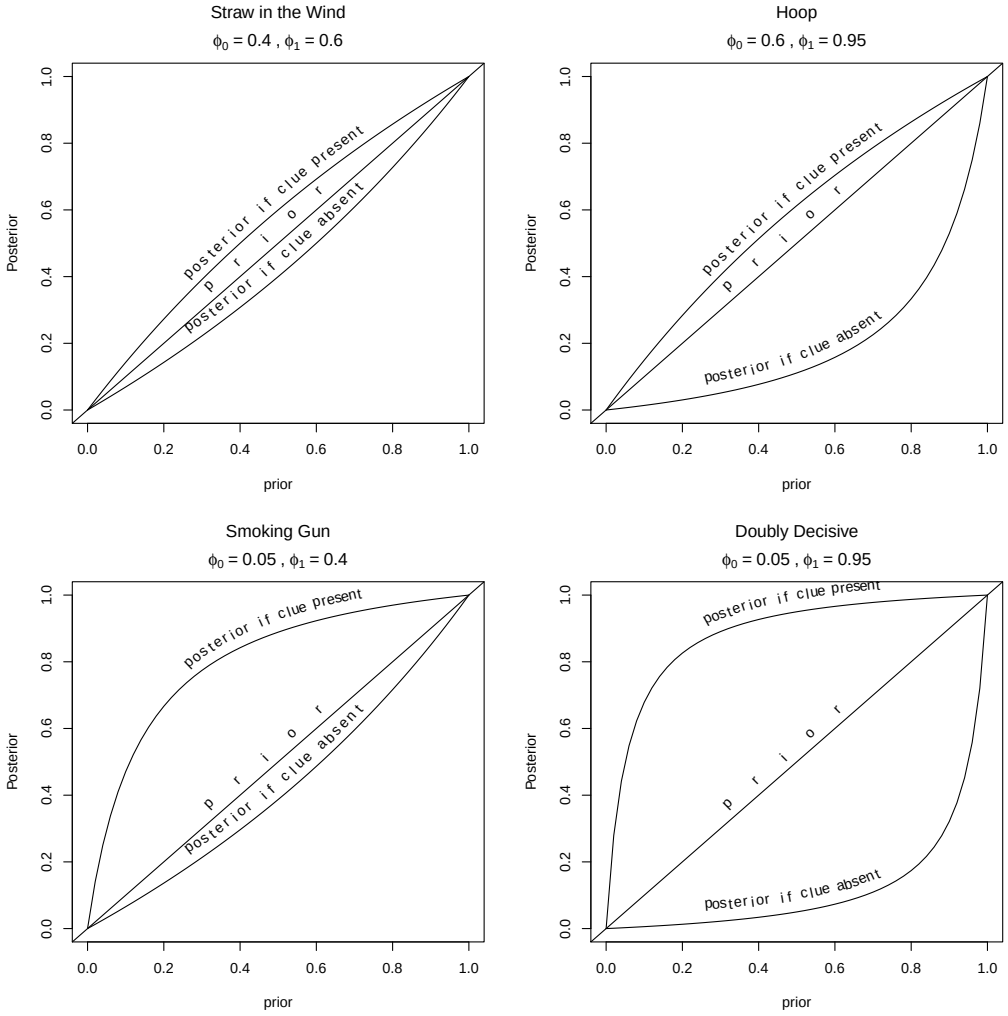

Figure 5.3: Figure shows how the learning from different types of tests depends on priors regarding the proposition. A smoking gun test has the greatest impact on beliefs when priors are middling low and the clue is observed; a ''hoop test" has the greatest effect when priors are middling high and the clue is not observed.

hoop test is failed. But suppose that the existing literature has given us a very high level of confidence that the hypothesis *is* right. This high prior confidence, sometimes referred to as a "base rate," is equivalent to believing that the causal effect exists in a very high proportion of cases. Thus, while any given case with a causal effect has only a 0.1 chance of not generating the clue, the high base rate means that the vast majority of cases that we observe without the clue will nonetheless be cases with causal effects. Thus, the failure of even a strong hoop test, involving a highly certain prediction, should only marginally reduce our confidence in a hypothesis that we strongly expect to be true.

A similar line of reasoning applies to smoking gun tests involving hypotheses that prior evidence suggests are very unlikely to be true. Innocent people may be very unlikely to be seen holding smoking guns after a murder. But if a very high proportion of people observed are known to be innocent, then a very high proportion of those holding smoking guns will in fact be innocent—and a smoking-gun clue will be far from decisive.

We emphasize two respects in which these implications depart from common intuitions. First, we cannot make *general* statements about how decisive different categories of test, in Van Evera's framework, will be. It is commonly stated that hoop tests are devastating to a theory when they are failed, while smoking gun tests provide powerful evidence in favor of a hypothesis. But, in fact the amount learned depends not just on features of the clues but also on prior beliefs.

Second, although scholars frequently treat evidence that goes against the grain of the existing literature as especially enlightening, in the Bayesian framework the contribution of such evidence may sometimes be modest, precisely because received wisdom carries weight. Thus, although the discovery of *disconfirming* evidence—an observation thought to be strongly inconsistent with the hypothesis—for a hypothesis commonly believed to be true is more informative (has a larger impact on beliefs) than *confirming* evidence, this does not mean that we learn more than we would have if the prior were weaker. % But it is not true as a general proposition that we learn more the bigger the "surprise" a piece of evidence is. %The effect of disconfirming evidence on a hypothesis about which we are highly confident will be *smaller* than it would be for a hypothesis about which we are only somewhat confident. When it comes to very strong hypotheses, the "discovery"

of disconfirming evidence is very likely to be a false negative; likewise, the discovery of supporting evidence for a very implausible hypothesis is very likely to be a false positive. The Bayesian approach takes account of these features naturally.[4]

5.3.2 Simultaneous, joint updating

When we update we often update over multiple quantities. When we see a smoking gun, for instance, we might update our beliefs that the butler did it, but we might also update our beliefs about how likely we are to see smoking guns – maybe they are not so rare as we thought!

Intuitively you might think of this updating as happening sequentially – first of all you update over the general proposition, then you update over the particular claim. But in fact you update over both quantities at once.

Here we elaborate on the intuition of fully Bayesian process tracing, in which updating occurs over both causal type (j) and beliefs about the probabilities with which clues are observed for each type (ϕ values). The illustration in the text makes clear how updating over type occurs, given beliefs about ϕ values. But how does updating over ϕ occur?

Suppose that we observe a case with values $X = 1, Y = 1$. We begin by defining a prior probability distribution over each parameter. Suppose that we establish a prior categorical distribution reflecting uncertainty over whether the case is a b type (e.g., setting a probability of 0.5 that it is a b and 0.5 that is a d type). We also start with priors on ϕ_b and ϕ_d. For concreteness, suppose that we are certain that the clue is unlikely for a d type ($\phi_d = .1$), but we are very uncertain about ϕ_b; in particular, we have a uniform prior distribution over $[0, 1]$ for ϕ_b. Note that, even though we are very uncertain about ϕ_b, the clue still has probative value, arising from the fact that the expected value of ϕ_b is higher than that of ϕ_d.

Suppose that we then look for the clue in the case and observe it. This observation shifts posterior weight away from a belief that the case is a b. See Figure **??** for an illustration. Yet it *simultaneously* shifts weight toward

[4] We note, however, that one common intuition—that little is learned from disconfirming evidence on a low-plausibility hypothesis or from confirming evidence on a high-plausibility one—*is* correct.

a higher value for ϕ_b and a lower value for ϕ_d. The reason is that the observed clue has a relatively high likelihood *both* for combinations of parameter values in which the case is a d and ϕ_b is low *and* for combinations in which the case is a b and ϕ_b is *high* (or, equivalently, in this example, where ϕ_d is low). The marginal posterior distribution of ϕ_b will thus be shifted upward relative to its prior marginal distribution. The joint posterior distribution will also reflect a dependency between the probability that the case is a b vs. a d, on the one hand, and ϕ_b and ϕ_d on the other.

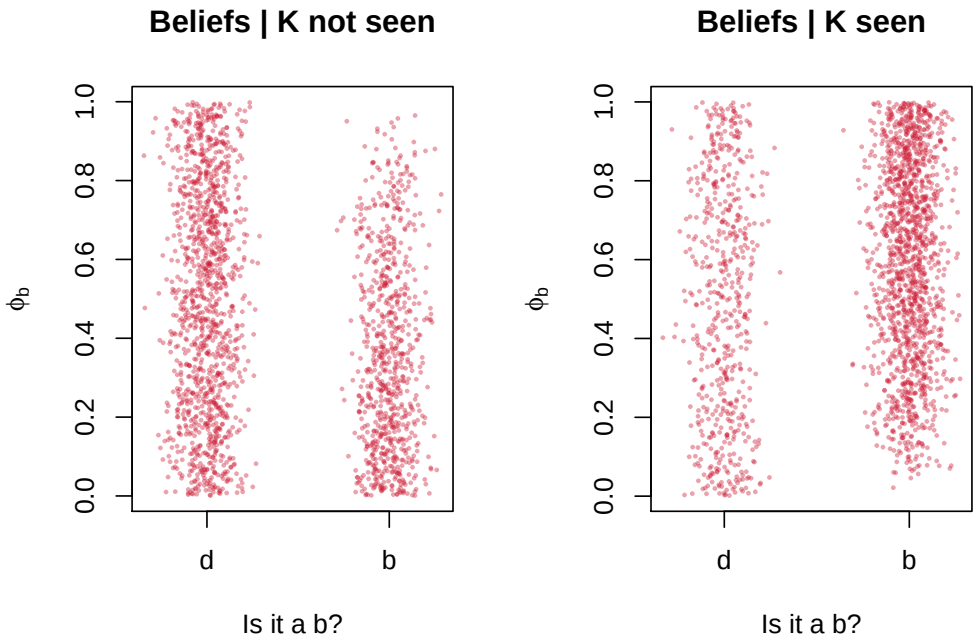

Figure 5.4: Joint posteriors distribution on whether a case is a b or d and on the probability of seeing a clue for a b type (ϕ_b).

5.3.3 Posteriors are independent of the ordering of data

We often think of learning as a process in which we start off with some set of beliefs—our priors, we gather data, D_1, and update our beliefs, forming a posterior; we then observe new data and we update again, forming a new

posterior, having treated the previous posterior as a new prior. In such cases it might seem natural that it matters which data we saw first and which later.

For instance EXAMPLE

In fact though, Bayesian updating is deaf to ordering. If we learn first that the card is a face card and second that it is black, our posteriors that it is a Jack of Spades go from 1 in 52 to 1 in 12 to 1 in 6. If we learn first that the card is black and second that it is a face card, our posteriors that it is a Jack of Spades go from 1 in 52 to 1 in 26 to 1 in 6. We end up in the same places in both cases. And we would ave had the same conclusion if we learned in one go that the card is a black face card.

The math of this is easy enough. Our posterior given two sets of data D_1, D_2 can be written:

$$p(\theta|D_1, D_2) = \frac{p(\theta, D_1, D_2)}{p(D_1, D_2)} = \frac{p(\theta, D_1|D_2)p(D_2)}{p(D_1|D_2)p(D_2)} = \frac{p(\theta, D_1|D_2)}{p(D_1|D_2)}$$

or, equivalently:

$$p(\theta|D_1, D_2) = \frac{p(\theta, D_1, D_2)}{p(D_1, D_2)} = \frac{p(\theta, D_2|D_1)p(D_1)}{p(D_2|D_1)p(D_1)} = \frac{p(\theta, D_2|D_1)}{p(D_2|D_1)}$$

In other words our posteriors given both D_1 and D_2 can be thought of as the result of updating on D_2 given we already know D_1 or the result of updating on D_1 given we already know D_2.

This fact will be useful in applications. In practice we might assume that we have beliefs based on background data D_1, for example regarding general relations between X and Y and a flat prior, and we then update again with new data on K. Rather than updating twice, the fact that updating is invariant to order means that we can assume a flat prior and update once given data on X, Y, and K.

Part II

Model-Based Causal Inference

Chapter 6

Process Tracing with Causal Models

We connect the literature on causal models to qualitative inference strategies used in process tracing. We provide a procedure for inference on case level queries from causal models. In addition we extract a set of implications for process tracing. We show how a key result from the causal models literature provides a condition for when clues may be (or certainly will not be) informative.

6.1 Process tracing and causal models

This chapter demonstrates how we can use causal models to conduct confirmatory process tracing: that is, to draw causal inferences about a single case from case-level data.

6.1.1 The intuition

We first walk through the basic intuition and then provide a more formal account.

When we undertake process tracing, we seek to answer a causal query about a given case. The key insight driving our approach is that **the inference about a causal estimand for a case is a claim about what causal types are both likely ex ante (given prior knowledge) and consistent with the data**.[1]

The estimand of interest can be a statement about any number of case-level causal features, including a case-level causal effect, the pathway through which an effect operates, an actual cause, or causal attribution. We will use observations from the case itself to address this query. We do so via a procedure in which we first encode prior knowledge in the form of a causal model, use data to learn about features of the model, and then take what we have learned about the model and map it into our query.

Given a causal model, we form posteriors over estimands as follows:

1. **Specify all causal types**. A causal type, recall, specifies the values that a unit is expected to take, absent any interventions, but also the values it would take given some interventions on some variables. Examples of types might be:

- Type 1: $(X = 1)$ *and* $(Y = 1$ if $X = 1$, $Y = 0$ if $X = 0)$.
- Type 2: $(X = 0)$ *and* $(Y = 1$ if $X = 1$, $Y = 0$ if $X = 0)$.
- Type 3: $(X = 1)$ *and* $(Y = 1$ if $X = 1$, $Y = 1$ if $X = 0)$.

2. **Specify priors over causal types.** Report how likely you think it is that a given unit is of a particular causal type. In the simplest case one might place 0 weight on some causal types (that might be ruled out by theory, for example) and equal weight on the others.

3. **Specify the estimand in terms of causal types.** For instance the estimand "Y responds positively to X" can be thought of as a collection of causal types: Q={Type 1, Type 2}.[2]

4. **Specify the set of causal types that are consistent with the data.** For instance if we observe $X = 1, Y = 1$ we might specify the data-consistent set as {Type 1, Type 3.}.

[1] This differs from the task for mixed methods that we will address in Chapter 8 as these concern claims about the distribution of causal types in populations.

[2] More generally an estimand might be a function of the distribution of causal types.

5. **Update.** Updating is done then by adding up the prior probabilities on all causal types that are consistent with both the data and the estimand, and dividing this by the sum of prior probabilities on all causal types that are consistent with the data (whether or not they are consistent with the estimand).

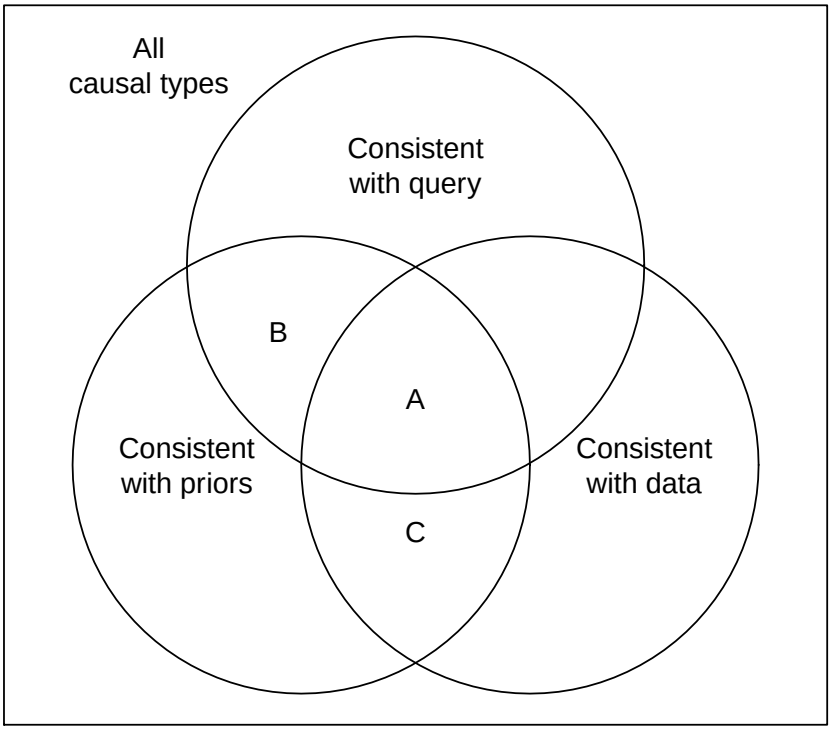

Figure 6.1: Logic of simple updating on arbitrary estimands.

This process is represented graphically with Figure **??**, where we can think of probabilities as proportionate to areas. Our causal model defines the causal

type space. We then proceed by a process of elimination. Only some of the causal types in the model are consistent with prior knowledge. Only some are consistent with the data that we observe. Finally, any query itself maps onto a subset of the possible causal types. The causal types that remain in contention once we have observed the evidence are those at the intersection of consistency with priors and consistency with the data. A represents those types that are *also* consistent with a given answer to the query (say, X has a positive effect on Y).

Thus, our belief about the query before we have seen the data is the probability of all causal types consistent with our priors and with the query $(A + B)$ as a proportion of all types consistent with our priors. Once we have seen the data, we have reduced the permissible types to $A + C$. Our posterior belief on the query is, then, the probabilities of those remaining types that are consistent with the query as a share of the probabilities of *all* remaining types, or $A/(A + C)$.

What we are doing here is straightforward: assessing causal possibilities for their compatibility with both the evidence at hand and our prior knowledge of how the world works. The formalization that we will present ensures that prior knowledge and evidence are all recorded explicitly while forcing logical consistency on the inferences that emerge from them.

6.1.2 A formalization of the general approach

More formally, the general approach to inference draws on the components we outlined in chapters 2 to 4: graphical causal models (DAGs), nodal and causal types, and priors. We now show how these elements formally interact with data to generate causal inferences. We continue to focus on a situation with binary variables, though suggest later in the chapter how this can be extended. Though we walk through the procedure for simple models, the approach outlined here can be applies to *any* causal model with binary variables and to any estimands defined over the model.

The process tracing procedure operates as follows:

A DAG. We begin with a DAG, or graphical causal model. As we know, a DAG identifies a set of variables and describes the parent-child relations between them, indicating for each variable which other variables are its direct (possible) causes. These relationship, in turn, tell us which (non-descendant)

variables a given variable is *not* independent of given the other variables in the model.

Nodal types. Once we have specified a DAG, we have defined the full set of possible nodal types: the types defining the value that a variable will take on given the values of its parents, which we have denoted with θ values. At each node, the range and number of nodal types is defined by the number of parents that that node has and the number of values the variables can take on. For instance, assuming all variables to be binary, if Y has parents X and W (so $k = 2$), then there are $2^{(2^2)} = 16$) possible causal types for the Y node. There are 2^2 possible combinations of values that two binary causal variables can take on—-$(X = 0, W = 0), (X = 0, W = 1), (X = 1, W = 0), (X = 1, W = 1)$—which implies four possible causal conditions over which Y's possible responses must be defined. For instance, as we have seen, with two causal variables, we can have θ^Y_{0000}, where Y is always 0; θ^Y_{0001}, where Y is 0 unless both X and W are 1; and so on.[3] To get the total number of nodal types, we simply raise 2 (since Y is binary) to the number of causal conditions (4), giving the number of possible patterns of Y values that could be generated across these four conditions (16). (The full set of nodal types for two causal variables in a binary setup is given in **??**.)[4]

All variables in a model have nodal types defining the value they take on given the value of their parents, including those variables without substantive parents. Suppose that X and W, in this model, have no substantively defined parents. We nonetheless define a nodal type for each of them, which simply captures their exogenous assignment to some value. With X binary, for instance, there are two nodal types, θ^X_0, where X is set to 0, and θ^X_1, where

[3]These nodal types can require many indices–2^k for a node with k parents—and the rule we follow is that the ith subscript indicates the value the node takes when parent $j \in 1, 2, ..., k$ take values $\mod (floor((i-1)/(2^{j-1})), 2)$ For instance for `Y0111` the first index means that Y takes the value 0 where both parents are 0, in all other cases it takes value 1.

[4]More generally, let us say that any node j can take on r_j possible values and has parents belonging to set PA_j and that each parent, $i \in PA_j$, can take on r_i values. Then the number of nodal types for node j is equal to $r_j^{\prod_{i \in PA_j} r_i}$. Informally, the exponent in this expression simply multiplies by one another the number of values that each of j's parents can take on. This product tells us the number of causal conditions across which j's responses must be defined. We then raise the number of values that j can take on to the power of the number of causal conditions. With all variables binary, this expression translates to $2^{(2^k)}$ nodal types for a node with k parents.

X is set to 1.

Unit causal types. We will want to be able to conceive not just of types for individual nodes but of the full collection of nodal types across all nodes in a model. We refer to a unit's full set of nodal types as its *unit causal type* — or, more simply, unit type — which we represent as θ. A unit type is simply a listing that contains one nodal type for each node in the model. For instance, with a model with variable X, W, and Y, each unit has a *causal* type composed of its *nodal* types on each of the three nodes.[5] Thus, one causal type in this model could be $\theta = (\theta^X = \theta^X_1, \theta^W = \theta^W_1, \theta^Y = \theta^Y_{1101})$. Another could be $\theta = (\theta^X = \theta^X_0, \theta^W = \theta^W_1, \theta^Y = \theta^Y_{0001})$. And so on.

We show the mapping between nodal and causal types, for a simply $X \rightarrow Y$ model, in Table **??**. The column headings represent the 8 permissible causal types, each expressed simply as a concatenated strings of nodal types. The row headings represent the nodal types. In each interior cell, a 1 or 0 indicates whether or not a given nodal type is a component of a given causal type. As can be seen, each causal type has two nodal types that are its components since there are two nodes in this model. Each X-nodal type is part of four causal types since it can be combined with four different Y-nodal types, while each Y-nodal type is part of two causal types since it can be combined with two X-nodal types.

Table 6.1: . A mapping between nodal types and causal types for a simple $X \rightarrow Y$ model.

Causal Types $\rightarrow$	$\theta^X_0.\theta^Y_{00}$	$\theta^X_1.\theta^Y_{00}$	$\theta^X_0.\theta^Y_{10}$	$\theta^X_1.\theta^Y_{10}$	$\theta^X_0.\theta^Y_{01}$	$\theta^X_1.\theta^Y_{01}$	$\theta^X_0.\theta^Y_{11}$	$\theta^X_1.\theta^Y_{11}$
Nodal types $\downarrow$								
θ^X_0	1	0	1	0	1	0	1	0
θ^X_1	0	1	0	1	0	1	0	1
θ^Y_{00}	1	1	0	0	0	0	0	0
θ^Y_{10}	0	0	1	1	0	0	0	0

[5] A model in which each node j has k_j parents has $\prod_j 2^{(2^{k_j})}$ causal types that uniquely determine what data will be observed for a type under all possible interventions on its exogenous nodes.

Causal Types $\rightarrow$	$\theta_0^X.\theta_{00}^Y$	$\theta_1^X.\theta_{00}^Y$	$\theta_0^X.\theta_{10}^Y$	$\theta_1^X.\theta_{10}^Y$	$\theta_0^X.\theta_{01}^Y$	$\theta_1^X.\theta_{01}^Y$	$\theta_0^X.\theta_{11}^Y$	$\theta_1^X.\theta_{11}^Y$
θ_{01}^Y	0	0	0	0	1	1	0	0
θ_{11}^Y	0	0	0	0	0	0	1	1

Priors: Our background beliefs about a causal domain will usually consist of more than just beliefs about which variables have causal connections; they will also typically contain beliefs about what *kinds* of effects operate between variables. That is, they will contain beliefs about which types are possible or, more generally, are more or less common in the world. We express these beliefs over causal effects as either restrictions on nodal types or as probability distributions over the nodal types.

In general, when doing process tracing in this framework, we think of a given case of interest – the one we are studying and seek to learn about – as being drawn at random from a population. Thus, our prior beliefs about a *single* case – before we do the process tracing – are really beliefs about that population. So, for instance, our prior belief about the probability that inequality has a positive effect on democratization in Mexico in 1999 is our belief about how commonly inequality has a positive effect on democratization in the population of cases that are "like" Mexico in 1999.[6]

We let λ^j denote our belief about the population distribution of nodal types at node j. A λ^j is simply a vector of proportions, one for each possible nodal type, with the proportions adding up to 1. So, for instance, λ^Y for our current example would be a vector with four values, each of which expresses a proportion for one of the four nodal types at Y. So we might have $\lambda_{01}^Y = 0.1$, $\lambda_{11}^Y = 0.05$, and so on – with the λ^Y values summing to 1 because these values are defined over the full set of possible nodal types for Y.

We can, in turn, use these population parameters – these beliefs about nodal-type proportions in the population – to create prior probabilities over the *causal* type for the case at hand. Since causal types are merely combinations

[6]The reference population for a case is defined based on whatever we already know about the case. Thus, for instance, if we already know that the case has $Y = 1$ before we begin process tracing, then the relevant population for the formation of prior beliefs is all cases in which $Y = 1$.

of nodal types, and our case has been drawn at random from the population, we can take a set of posited proprtions of nodal types in the population and readily calculate the probability that our case is of any given causal type. To do so, we need to join together λ's across the nodes in a model.

Let us first see how this works in a situation in which we assume that the nodal types are independent of one another. We can think of this as a situation in which there is no confounding that is not captured in the graph – no variable missing from the model that is a common ancestor of multiple nodes in the model. Here, our beliefs over causal types are simply the product of our beliefs over the component nodal types (since the joint probability of independent events is simply the product of their individual probabilities). For instance, one causal type might be "a unit in which $X = 1$ and in which $Y = 1$ no matter what value X takes." In this case the probability that a case is of this causal type might be written $\Pr(\theta^X = \theta^X_1) \Pr(\theta^Y = \theta^Y_{11}) = \lambda^X_1 \lambda^Y_{11}$.

The simplest way in which we can express beliefs about the differential probabilities of different causal possibilities is by *eliminating* nodal types that we do not believe to be possible—setting their parameter values to 0. Suppose, for instance, that we are examining the effect of ethnic diversity on civil war in a case. We might not know whether ethnic diversity causes civil war in this case, but we might have sufficient background knowledge to believe that ethnic diversity never has a *negative* effect on civil war: it never prevents a civil war from happening that would have happened in the absence of ethnic diversity. We would thus want to set the parameter value for a negative causal effect to 0. If we then know nothing about the relative frequencies of the three remaining nodal types for Y, we may (following the principle of indifference), frequency of positive effects, null effects with civil war destined to happen, and null effects with civil war never going to happen, assigning a weight of $\frac{1}{3}$ to each of them.

In a situation of unobserved confounding, our beliefs over causal types are still well defined, though they are no longer the simple product of beliefs over nodal types. Let us imagine for instance, in a simple $X \rightarrow Y$ model, that we believe that some unobserved factor both makes cases more likely to have $X = 1$ and makes it more likely that X has a positive effect on Y. This is the same as saying that the probability that $\theta^X = \theta^X_1$ is positively correlated with the probability that $\theta^Y = \theta^Y_{01}$. Now, our probability that *both* $X = 1$ and X has a positive effect must be calculated using the joint

probability formula, $\Pr(A, B) = \Pr(A)\Pr(B|A)$.[7] Thus, $\Pr(\theta^Y = \theta^Y_{01}, \theta^X = \theta^X_1) = \Pr(\theta^Y = \theta^Y_{01})\Pr(\theta^X = \theta^X_1|\theta^Y = \theta^Y_{01})$. To form priors over causal types in this situation, we need to posit beliefs about a set of more complex, conditional proportions for X's type. Specifically, we need to posit, *for those cases* with a positive effect of X on Y, what proportion are "assigned" to $X = 1$; and, separately, what proportion are assigned to $X = 1$ among those cases *without* a positive effect of X on Y.

These conditional proportions may, of course, be difficult for the researcher to form beliefs about. Forming a belief about them amounts to saying that we do not know what generates confounding, but we know the correlations it generates in the data. We may wonder how often we will be in that epistemological position. An alternative way to parse the problem, then, is to *model* the confounding by including the confounder (say, Z) as a new node in the graph. In the above example, Z would point into both X and Y. We would then posit population proportions for a set of nodal types for X – representing X's possible responses to Z – and for Y – representing Y's possible responses to both X and Z. We may find it easier to reason and form beliefs about these more complex nodal types than about the conditional proportions involved in unobserved confounding. The two approaches work out to be analytically equivalent given equivalent underlying beliefs, so the choice between them will be a matter of researcher preference.[8]

Importantly, in process tracing, we are focused on drawing case-level inferences and, as such, we treat the population-level parameters as given and fixed. In general, these parameters derive from our beliefs about how the world works, and those beliefs will typically be uncertain. The key point, however, is that in process tracing, the population parameters serve as an *input* into the analysis, conditioning our inferences from the evidence; but we do not *update* on these population-level beliefs once we see the data from a single case. Importantly, as we show later in the book, we *do* update on population-level inferences in the more general setup that we introduce in

[7] In words, the probability of A and B occurring is equal to the probability of A occurring times the probability of B occurring *given* that A occurs.

[8] As we will see later in the book, another approach is to gather data on additional cases. When analyzing multiple cases, we can set up our priors to allow for the possibility of unobserved confounding and then, potentially, learn about that confounding from the data. This is not possible under our procedure for single-case process tracing, where we treat the population parameters as given and fixed.

Chapter **??** for analyzing mixed data in multiple cases. We also show in Chapter **??** how we can test the sensitivity of conclusions to the values at which we set population parameters. Interestingly, as we also show, process-tracing inferences, including uncertainty about conclusions, are unaffected by the level of uncertainty we might have about population parameters; we thus do not specify this uncertainty for the purposes of process tracing.

The relationship between causal types, nodal types, and the correlation among nodal types is captured in what we call a *parameter matrix.* We show a parameter matrix for a simple $X \rightarrow Y$ model with no unobserved confounding in Table **??**. Here each column label (except the last) represents the probability that a case is of a given causal type. Each row label represents a population-level parameter: a belief about the proportions of different nodal types in the population. We indicate a set of possible parameter values in the final column.

Table 6.2: . A mapping between nodal types and causal types for a simple $X \rightarrow Y$ model (with no unobserved confounding).

Causal types $\rightarrow$	$\theta^X_0, \theta^Y_{00}$	$\theta^X_1, \theta^Y_{00}$	$\theta^X_0, \theta^Y_{10}$	$\theta^X_1, \theta^Y_{10}$	$\theta^X_0, \theta^Y_{01}$	$\theta^X_1, \theta^Y_{01}$	$\theta^X_0, \theta^Y_{11}$	$\theta^X_1, \theta^Y_{11}$	Parameter values (population proportions)
Population parameters $\downarrow$									
λ^X_1	0	1	0	1	0	1	0	1	0.5
λ^X_0	1	0	1	0	1	0	1	0	0.5
λ^Y_{00}	1	1	0	0	0	0	0	0	0.2
λ^Y_{10}	0	0	1	1	0	0	0	0	0.2
λ^Y_{01}	0	0	0	0	1	1	0	0	0.4
λ^Y_{11}	0	0	0	0	0	0	1	1	0.2

To start with the first two rows, these represent the population proportions of each of X's nodal types. For instance, λ_0^X is our belief about the proportion of cases in the population that are of nodal type θ_0^X. The first row, λ_1^X, represents our belief about the inverse: the proportion of cases in the population of type θ_1^X. We posit beliefs about these parameters in the final column, indicating that we think that half of cases in the population are "assigned" to $X = 0$ and half to $X = 1$. Note that, since there are only two possible nodal types for X, and their proportions must sum to 1, there is actually just one degree of freedom here: once we've specified one of these parameter values, the other is defined as well.

The last four rows represent the proportion of cases in the population with different Y-nodal types: in order, the proportion in which X has no effect on Y, with Y fixed at 0; the proportion in which X has a negative effect; the proportion in which X has a positive effect; and the proportion in which X has no effect, with Y fixed at 1. Again, in the last column, we provide possible values for these proportions, the four of which must also sum to 1. Here we are stating that positive $X \rightarrow Y$ effects are twice as common in the population as the other three nodal types, which we set at equal prevalence.

The interior cells indicate whether a given population parameter enters into the prior probability of a given causal type. Thus, for instance, to calculate the prior probability of the causal type $\theta_1^X, \theta_{10}^Y$, we need to multiply the two parameters values corresponding to the 1's in this causal type's column: λ_1^X by λ_{10}^Y. Given the parameter values we have assigned for this example, then, the prior on this causal type is simply $0.5 \times 0.2 = 0.1$.

The prior probability that a case is of a given causal type thus comes directly from our beliefs about how nodal types are distributed in the population. All we know before we study a case is whatever we know about cases "like" it in general. It is then these causal-type probabilities – which represent probabilities that a *given case* is of a particular causal type – that we will update on once we see the data for this case.

We show the somewhat more complex situation of unobserved confounding in Table **??**. It is the first four rows that allow for unobserved confounding—the correlations across types. In a potential outcomes framework, we could think of these rows as capturing differential "assignment propensities" for X. Here, we allow for different probabilities of X's type being θ_1^X depending on what Y's type is. Thus, $\lambda_0^X | \theta^Y = \theta_{01}^Y$ is the proportion of θ_0^X types among cases

with θ^Y_{01} type: put differently, it is the probability of X being assigned to 0 when X has a positive effect on Y. The second row represents the inverse proportion: the proportion of a θ^X_1 types among θ^Y_{01} types. The next two rows then capture the proportions of the X-types among all *other* Y-types (i.e., among those cases for which X does *not* have a positive effect on Y).

Unobserved confounding in this setup takes the form of a difference in the proportions of a given X type among different Y types. Thus, if $\lambda^X_1, |\theta^Y_{01}$ is not the same as $\lambda^X_1|\theta^Y \neq \theta^Y_{01}$, we have unobserved confounding. Imagine, for instance, if we are studying the effect of faster economic growth (X) on democratization (Y), and we believe that there is some unobserved factor that both makes some countries' economies grow more quickly and also makes economic growth more likely to have a positive effect on democratization. This belief amounts to a belief that the probability of a case being assigned to $X = 1$ is higher if Y's nodal type is θ^Y_{01} than if it is not. In other words, in terms of the rows in Table **??**, we believe here that $\lambda^X_1|\theta^Y = \theta^Y_{01}$ is greater than $\lambda^X_1|\theta^Y \neq \theta^Y_{01}$. To illustrate, we provide parameter values along these lines in the final column.

Again, however, a researcher might prefer to specify the confounder (say, Z) as a node in the model. The rows in the parameter matrix would then be a set of population parameters defined as proportions of *un*conditional nodal types, with four X-types representing possible responses to Z, and 16 Y types, representing Y's possible responses to X and Z.

Table 6.3: . A mapping between nodal types and causal types for a simple $X \rightarrow Y$ model *with* unobserved confounding.

Causal Types $\rightarrow$	$\theta^X_0, \theta^Y_{00}$	$\theta^X_1, \theta^Y_{00}$	$\theta^X_0, \theta^Y_{10}$	$\theta^X_1, \theta^Y_{10}$	$\theta^X_0, \theta^Y_{01}$	$\theta^X_1, \theta^Y_{01}$	$\theta^X_0, \theta^Y_{11}$	$\theta^X_1, \theta^Y_{11}$
Population parameters $\downarrow$								
$\lambda^X_0\|\theta^Y = \theta^Y_{01}$	0	0	0	0	1	0	0	0

Causal Types $\rightarrow$	$\theta_0^X, \theta_{00}^Y$	$\theta_1^X, \theta_{00}^Y$	$\theta_0^X, \theta_{10}^Y$	$\theta_1^X, \theta_{10}^Y$	$\theta_0^X, \theta_{01}^Y$	$\theta_1^X, \theta_{01}^Y$	$\theta_0^X, \theta_{11}^Y$	$\theta_1^X, \theta_{11}^Y$
$\lambda_1^X \vert \theta^Y = \theta_{01}^Y$	0	0	0	0	0	1	0	0
$\lambda_0^X \vert \theta^Y \neq \theta_{01}^Y$	1	0	1	0	0	0	1	0
$\lambda_1^X \vert \theta^Y \neq \theta_{01}^Y$	0	1	0	1	0	0	0	1
λ_{00}^Y	1	1	0	0	0	0	0	0
λ_{10}^Y	0	0	1	1	0	0	0	0
λ_{01}^Y	0	0	0	0	1	1	0	0
λ_{11}^Y	0	0	0	0	0	0	1	1

One special kind of prior that we might wish to set is to disallow a particular (conditional) type altogether. For instance, if studying the effect of we may believe that

Possible data types. A *data type* is a particular pattern of data that we could potentially observe for a given case. More specifically, a data type is a set of values, one for each node in a model. For instance, in our X, W, Y setup, $X = 1, W = 0, Y = 0$ would be one data type.

Importantly, each possible causal type *maps into a single data type.* One intuitive way to think about why this is the case is that a causal type tells us (a) the values to which all exogenous variables in a model are assigned and (b) how all endogenous variables respond to their parents. Given these two components, only one set of node values is possible. For example, causal type $\theta = (\theta^X = \theta_1^X, \theta^W = \theta_0^W, \theta^Y = \theta_{0100}^Y)$ imples data $X = 1, W = 0, Y = 1$. There is no other set of data that can be generated by this causal type.

Equally importantly, however, *the mapping from causal types to data types is not one-to-one.* More than one causal type can generate the same case-level data pattern. For instance, the causal type $\theta = (\theta^X = \theta_1^X, \theta^W = \theta_0^W, \theta^Y = \theta_{1101}^Y)$ will *also* generate the data type, $X = 1, W = 0, Y = 1$. Thus, observing this data type leaves us with ambiguity about the causal type by which it was generated.

A full mapping between causal types and data types can be summarized by an "ambiguity matrix." In Table **??**, we provide an example of such a matrix, derived directly from the parameter matrix in Table **??**. Here, the rows represent causal types and the columns (except for the last) represent data types. The notation for data types is straightforward, with for instance $X0Y0$ meaning that $X = 0, Y = 0$ has been observed. In the interior cells, the 1's and 0's indicate whether or not a given data type could arise from a given causal type. We can readily see here that each causal type can generate only one data type.

We can also see the ambiguity of the data, however, since each data type can be generated by two causal types. For instance, if we observe $X = 1, Y = 1$, we know that the case is either of causal type $\theta_1^X, \theta_{01}^Y$ or of causal type $\theta_1^X, \theta_{11}^Y$ – but do not know which.

Table 6.4: . An ambiguity matrix, mapping from data types to causal types for a simple $X \rightarrow Y$ model.

Data types $\rightarrow$	X0Y0	X1Y0	X0Y1	X1Y1	Priors on causal types
Causal types $\downarrow$					
$\theta_0^X, \theta_{00}^Y$	1	0	0	0	0.1
$\theta_1^X, \theta_{00}^Y$	0	1	0	0	0.1
$\theta_0^X, \theta_{10}^Y$	0	0	1	0	0.1
$\theta_1^X, \theta_{10}^Y$	0	1	0	0	0.1
$\theta_0^X, \theta_{01}^Y$	1	0	0	0	0.2
$\theta_1^X, \theta_{01}^Y$	0	0	0	1	0.2
$\theta_0^X, \theta_{11}^Y$	0	0	1	0	0.1
$\theta_1^X, \theta_{11}^Y$	0	0	0	1	0.1

In the last column, we provide prior probabilities for each of the causal types. These have been calculated directly from the parameter matrix (Table **??**). To see how the calculation works, start with a causal type in the parameter matrix – say, $\theta_0^X, \theta_{01}^Y$. We go down that causal type's column and select the rows with 1's, representing the parameters for the included nodal types, λ_0^X and λ_{01}^Y. As we want the joint probability of these two nodal types (and a parameter matrix is constructed such that the rows represent independent

events),[9] we simply multiply together the values for these included parameters: $0.5 \times 0.4 = 0.2$. As noted, our prior belief about whether the case at hand is of a given causal type is a straightforward function of our beliefs about how prevalent each of the component nodal types is in the population.

As models get more complex, the numbers of causal and data types simply multiply. In Table **??**, we show the ambiguity matrix for a simple mediation model ($X \rightarrow M \rightarrow Y$). Here, the causal types are combinations of three nodal types, one for each variable in the model. Similarly, the data types have three elements, one for each variable. We now have 8 data types and 32 causal types.

Table 6.5: . An ambiguity matrix, mapping from data types to causal types for a simpe mediation model, $X \rightarrow M \rightarrow Y$.

Data types →	X0M0Y0	X0M0Y1	X0M1Y0	X0M1Y1	X1M0Y0	X1M0Y1	X1M1Y0	X1M1Y1	Priors on causal types
Causal types ↓									
$\theta_0^X, \theta_{00}^M, \theta_{00}^Y$	1	0	0	0	0	0	0	0	0.02
$\theta_1^X, \theta_{00}^M, \theta_{00}^Y$	0	1	0	0	0	0	0	0	0.02
$\theta_0^X, \theta_{10}^M, \theta_{00}^Y$	0	0	1	0	0	0	0	0	0.02
$\theta_1^X, \theta_{10}^M, \theta_{00}^Y$	0	1	0	0	0	0	0	0	0.02
$\theta_0^X, \theta_{01}^M, \theta_{00}^Y$	1	0	0	0	0	0	0	0	0.04
$\theta_1^X, \theta_{01}^M, \theta_{00}^Y$	0	0	0	1	0	0	0	0	0.04
$\theta_0^X, \theta_{11}^M, \theta_{00}^Y$	0	0	1	0	0	0	0	0	0.02
$\theta_1^X, \theta_{11}^M, \theta_{00}^Y$	0	0	0	1	0	0	0	0	0.02
$\theta_0^X, \theta_{00}^M, \theta_{10}^Y$	0	0	0	0	1	0	0	0	0.02
$\theta_1^X, \theta_{00}^M, \theta_{10}^Y$	0	0	0	0	0	1	0	0	0.02
$\theta_0^X, \theta_{10}^M, \theta_{10}^Y$	0	0	1	0	0	0	0	0	0.02
$\theta_1^X, \theta_{10}^M, \theta_{10}^Y$	0	0	0	0	0	1	0	0	0.02
$\theta_0^X, \theta_{01}^M, \theta_{10}^Y$	0	0	0	0	1	0	0	0	0.04
$\theta_1^X, \theta_{01}^M, \theta_{10}^Y$	0	0	0	1	0	0	0	0	0.04
$\theta_0^X, \theta_{11}^M, \theta_{10}^Y$	0	0	1	0	0	0	0	0	0.02
$\theta_1^X, \theta_{11}^M, \theta_{10}^Y$	0	0	0	1	0	0	0	0	0.02

[9] That is, when there is unobserved confounding, we express conditional proportions, making all of the proportions conditionally independent of one another.

Data types →	X0M0Y0	X0M0Y1	X0M1Y0	X0M1Y1	X1M0Y0	X1M0Y1	X1M1Y0	X1M1Y1	Priors on causal types
$\theta_0^X, \theta_{00}^M, \theta_{01}^Y$	1	0	0	0	0	0	0	0	0.04
$\theta_1^X, \theta_{00}^M, \theta_{01}^Y$	0	1	0	0	0	0	0	0	0.04
$\theta_0^X, \theta_{10}^M, \theta_{01}^Y$	0	0	0	0	0	0	1	0	0.04
$\theta_1^X, \theta_{10}^M, \theta_{00}^Y$	0	1	0	0	0	0	0	0	0.04
$\theta_0^X, \theta_{01}^M, \theta_{01}^Y$	1	0	0	0	0	0	0	0	0.08
$\theta_1^X, \theta_{01}^M, \theta_{01}^Y$	0	0	0	0	0	0	0	1	0.08
$\theta_0^X, \theta_{11}^M, \theta_{01}^Y$	0	0	0	0	0	0	1	0	0.04
$\theta_1^X, \theta_{11}^M, \theta_{01}^Y$	0	0	0	0	0	0	0	1	0.04
$\theta_0^X, \theta_{00}^M, \theta_{11}^Y$	0	0	0	0	1	0	0	0	0.02
$\theta_1^X, \theta_{00}^M, \theta_{11}^Y$	0	0	0	0	0	1	0	0	0.02
$\theta_0^X, \theta_{10}^M, \theta_{11}^Y$	0	0	0	0	0	0	1	0	0.02
$\theta_1^X, \theta_{10}^M, \theta_{11}^Y$	0	0	0	0	0	1	0	0	0.02
$\theta_0^X, \theta_{01}^M, \theta_{11}^Y$	0	0	0	0	1	0	0	0	0.04
$\theta_1^X, \theta_{01}^M, \theta_{11}^Y$	0	0	0	0	0	0	0	1	0.04
$\theta_0^X, \theta_{11}^M, \theta_{11}^Y$	0	0	0	0	0	0	1	0	0.02
$\theta_1^X, \theta_{11}^M, \theta_{11}^Y$	0	0	0	0	0	0	0	1	0.02

Again, the ambiguities arising from data patterns are apparent. For instance, if we observe $X = 1, M = 0, Y = 0$, we see that there are four causal types that could have generated this pattern. To unpack the situation a bit, these data tell us that $\theta^X = \theta_1^X$. But they do not tell us whether M's type is such that X has a negative effect on M (θ_{10}^M) or X has no effect with M fixed at 0 (θ_{00}^M). Similarly, we do not know whether M has a positive effect on Y (θ_{01}^Y) or no effect with Y fixed at 0 (θ_{00}^Y). This leaves four combinations of nodal types—four causal types—that are consistent with the data.

Our priors here derive from a set of parameter values, much like in the previous example, in which the X types are equally common (0.5 each); a positive effect of X on M is twice as common (0.4) as the other M types (all set to 0.2); and a positive effect of M on Y is twice as common (0.4) as all other Y types (all at 0.2). We can then easily see why we thus get priors on some causal types are higher than those on others: for instance, the two causal types with priors of 0.08 both have two positive effects (at the $X \rightarrow Y$ and $M \rightarrow Y$ stages) while the causal types with priors of 0.02 include no

positive effects at either stage.

Updating on types given the data. Once we observe actual data in a case, we can then update on the probabilities assigned to each causal type. The logic is simple. When we observe a set of data from a case, we place 0 probability on all causal types that could not have produced these data; we then scale up the probabilities on all causal types that could have.

We can see how this works within an ambiguity matrix. Let's return to the ambiguity matrix in Table **??**. We start out with a set of probability weights on all rows (causal types). Now, suppose that we observe the data $X = 1, Y = 1$, i.e., data type $X1Y1$. We then look down the $X1Y1$ column, and we know that all rows with a 0 in them represent causal types that *could not have* generated these data. These causal types are thus excluded. What is left are two rows: $\theta_1^X, \theta_{01}^Y$ and $\theta_1^X, \theta_{11}^Y$. Returning now to the probabilities, we put 0 weight on all of the excluded rows; and then we scale up the remaining probabilities so that they sum to 1 (preserving the ratio between them). The priors of 0.2 and 0.1 in the retained rows scale up to $\frac{2}{3}$ and $\frac{1}{3}$, which become our *posterior* probabilities on the causal types. We display an updated ambiguity matrix, with excluded data types and causal types removed, in Table **??**.

Before we see any data on the case at hand, then, we believe (based on our beliefs about the population to which the case belongs) that there is a 0.2 probability that the case is one in which X is assigned to 1 and has a positive effect on Y; and 0.1 probability that it's a case in which X gets assigned to 1 and has no effect on Y with Y fixed at 1. Seeing the $X = 1, Y = 1$ data, we now believe that there is a 0.667 probability that the case is of the former type, and a 0.333 probability that it is of the latter type.

Table 6.7: . An updated version of the ambiguity matrix in Table **??**, after observing $X = 1, Y = 1$ in a case.

Data types →	X1Y1	Priors on causal types	Posteriors on causal types
Causal types ↓			
$\theta_1^X, \theta_{01}^Y$	1	0.2	0.6667
$\theta_1^X, \theta_{11}^Y$	1	0.1	0.3333

Table 6.6: Ambiguity matrix for X -> M -> Y model. Rows are causal types, columns are data types. Last column shows possible priors over rows.

	X0M0Y0	X1M0Y0	X0M1Y0	X1M1Y0	X0M0Y1	X1M0Y1	X0M1Y1
X0M00Y00	1	0	0	0	0	0	0
X1M00Y00	0	1	0	0	0	0	0
X0M10Y00	0	0	1	0	0	0	0
X1M10Y00	0	1	0	0	0	0	0
X0M01Y00	1	0	0	0	0	0	0
X1M01Y00	0	0	0	1	0	0	0
X0M11Y00	0	0	1	0	0	0	0
X1M11Y00	0	0	0	1	0	0	0
X0M00Y10	0	0	0	0	1	0	0
X1M00Y10	0	0	0	0	0	1	0
X0M10Y10	0	0	1	0	0	0	0
X1M10Y10	0	0	0	0	0	1	0
X0M01Y10	0	0	0	0	1	0	0
X1M01Y10	0	0	0	1	0	0	0
X0M11Y10	0	0	1	0	0	0	0
X1M11Y10	0	0	0	1	0	0	0
X0M00Y01	1	0	0	0	0	0	0
X1M00Y01	0	1	0	0	0	0	0
X0M10Y01	0	0	0	0	0	0	1
X1M10Y01	0	1	0	0	0	0	0
X0M01Y01	1	0	0	0	0	0	0
X1M01Y01	0	0	0	0	0	0	0
X0M11Y01	0	0	0	0	0	0	1
X1M11Y01	0	0	0	0	0	0	0
X0M00Y11	0	0	0	0	1	0	0
X1M00Y11	0	0	0	0	0	1	0
X0M10Y11	0	0	0	0	0	0	1
X1M10Y11	0	0	0	0	0	1	0
X0M01Y11	0	0	0	0	1	0	0
X1M01Y11	0	0	0	0	0	0	0
X0M11Y11	0	0	0	0	0	0	1
X1M11Y11	0	0	0	0	0	0	0

We can also see how this works for our $X \rightarrow M \rightarrow Y$ model, and the ambiguity matrix in Table **??**. If we observe the data $X = 1, M = 0, Y = 0$, for instance, this exercise would yield the updated ambiguity matrix in Table **??**(tab:ambigmedupdate). Here we have eliminated all rows (causal types) with a 0 in the relevant data-type column ($X1M0Y0$) and formed the posteriors by scaling up the priors in the retained rows.

Table 6.8: . An updated version of the ambiguity matrix in Table **??**, after observing $X = 1, M = 0, Y = 0$ in a case.

Data types $\rightarrow$		Priors on X1M0Y0 causal types	Posteriors on causal types
Causal types $\downarrow$			
$\theta_1^X, \theta_{00}^M, \theta_{00}^Y$	1	0.02	0.1667
$\theta_1^X, \theta_{10}^M, \theta_{00}^Y$	1	0.02	0.1667
$\theta_1^X, \theta_{00}^M, \theta_{01}^Y$	1	0.04	0.3333
$\theta_1^X, \theta_{10}^M, \theta_{01}^Y$	1	0.04	0.3333

A notable feature of the logic of single-case process tracing is that the relative probabilities on the retained causal types never change. If we start out believing that causal type A is twice as likely as causal type B, and both A and B are retained once we see the data, then A will be twice as likely as B in our posteriors. All updating occurs by *eliminating* causal types from consideration and zeroing in on those that remain.

type	X1M0Y0	prior	posterior
X1M00Y00	1	0.02	0.1667
X1M10Y00	1	0.02	0.1667
X1M00Y01	1	0.04	0.3333
X1M10Y01	1	0.04	0.3333

A similar logic applies if partial data are observed: that is, if we do not collect data for all nodes in the model. The one difference is that, now, rather than reducing to one column we entertain the possibility of any data *type* consistent with the *observed data*. In general, more than one data type will be consistent with partial data. For instance, suppose that we observe $X = 1, Y = 0$ but do not observe M's value. These are data that are

consistent with both the data type $X1M0Y0$ and the data type $X1M1Y0$ (since the unobserved M could be either 0 or 1). We thus retain both of these data-type columns as well as all causal types consistent with *either* of these data types. This gives the updated ambiguity matrix in Table **??**. We note that, with these partial data, we are not able to update as strongly. For instance, for the causal type $\theta_1^X, \theta_{00}^M, \theta_{00}^Y$, instead of updating to a posterior probability of 0.1667, we update to a posterior of only 0.0833 – because there is a larger set of causal types with which these partial data are consistent.

Table 6.9: . An updated version of the ambiguity matrix in Table **??**, after observing partial data in case: $X = 1, Y = 0$, with M unobserved.

Data types →	X1M0Y0	X1M1Y0	Priors on causal types	Posteriors on causal types
Causal types ↓				
$\theta_1^X, \theta_{00}^M, \theta_{00}^Y$	1	0	0.02	0.0833
$\theta_1^X, \theta_{10}^M, \theta_{00}^Y$	1	0	0.02	0.0833
$\theta_1^X, \theta_{01}^M, \theta_{00}^Y$	0	1	0.04	0.1667
$\theta_1^X, \theta_{11}^M, \theta_{00}^Y$	0	1	0.02	0.0833
$\theta_1^X, \theta_{01}^M, \theta_{10}^Y$	0	1	0.04	0.1667
$\theta_1^X, \theta_{11}^M, \theta_{10}^Y$	0	1	0.02	0.0833
$\theta_1^X, \theta_{00}^M, \theta_{01}^Y$	1	0	0.04	0.1667
$\theta_1^X, \theta_{10}^M, \theta_{01}^Y$	1	0	0.04	0.1667

type	X1M0Y0	X1M1Y0	prior	posterior
X1M00Y00	1	0	0.02	0.0833
X1M10Y00	1	0	0.02	0.0833
X1M01Y00	0	1	0.04	0.1667
X1M11Y00	0	1	0.02	0.0833
X1M01Y10	0	1	0.04	0.1667
X1M11Y10	0	1	0.02	0.0833
X1M00Y01	1	0	0.04	0.1667
X1M10Y01	1	0	0.04	0.1667

Updating on estimands. We now have a posterior probability for each causal type for the case at hand. The causal question we are interested in

answering, our estimand, may not be about causal types *per se.* It is about an estimand that can be expressed as a *combination* of causal types.

For instance, suppose we are working with the model $X \rightarrow M \rightarrow Y$; and that our question is, "Did $X = 1$ cause $Y = 1$?". This question is asking both:

1. Does $X = 1$ in this case?
2. Does X have a positive effect on Y in this case?

The causal types that qualify are those, and only those, in which the answer to both is "yes."

Meeting condition (1) requires that $\theta^X = \theta^X_1$.

Meeting condition (2) requires that θ^M and θ^Y are such that X has an effect on M that yields a positive effect of X on Y. This could occur via a positive $X \rightarrow M$ effect linked to a positive $M \rightarrow Y$ effect or via a negative $X \rightarrow M$ effect linked to a negative $M \rightarrow Y$ effect.

Thus, the qualifying causal types in this model are:

- $\theta^X_1, \theta^M_{01}, \theta^Y_{01}$
- $\theta^X_1, \theta^M_{10}, \theta^Y_{10}$

Our *prior* on the estimand—what we believe before we collect data on the case at hand—is given simply by summing up the prior probabilities on each of the causal types that correspond to the estimand. Note that we must calculate the prior from the full ambiguity matrix, before excluding types for inconsistency with the data. Returning to the full ambiguity matrix in Table **??**, we see that the priors on these two types (given the population parameters assumed there) are 0.08 and 0.02, respectively, giving a prior for the estimand of 0.1.

The posterior on any estimand is, likewise, given by summing up the posterior probabilities on each of the causal types that correspond to the estimand, drawing of course from the updated ambiguity matrix. For instance, if we observe the data $X = 1, M = 1, Y = 1$, we update to the ambiguity matrix in Table **??**. Our posterior on the estimand, "Did $X = 1$ cause $Y = 1$?" is the sum of the posteriors on the above two causal types. Since $\theta^X_1, \theta^M_{10}, \theta^Y_{10}$ is excluded by the data, this just leaves the posterior on $\theta^X_1, \theta^M_{01}, \theta^Y_{01}$, 0.4444, which is the posterior belief on our estimand.

If we observe only the partial data, $X = 1, Y = 1$, then we update to the ambiguity matrix in Table **??**. Now both corresponding causal types are included, and we sum their posteriors to get the posterior on the estimand: $0.0769 + 0.3077 = 0.3846$.

type	X1M1Y1	prior	posterior
X1M01Y01	1	0.08	0.4444
X1M11Y01	1	0.04	0.2222
X1M01Y11	1	0.04	0.2222
X1M11Y11	1	0.02	0.1111

Table 6.10: . An updated version of the ambiguity matrix in Table **??**, after observing $X = 1, M = 1, Y = 1$ in a case.

Data types $\rightarrow$		Priors on X1M1Y1 causal types	Posteriors on causal types
Causal types $\downarrow$			
$\theta_1^X, \theta_{01}^M, \theta_{01}^Y$	1	0.08	0.4444
$\theta_1^X, \theta_{11}^M, \theta_{01}^Y$	1	0.04	0.2222
$\theta_1^X, \theta_{01}^M, \theta_{11}^Y$	1	0.04	0.2222
$\theta_1^X, \theta_{11}^M, \theta_{11}^Y$	1	0.02	0.1111

Table 6.11: . An updated version of the ambiguity matrix in Table **??**, after observing partial data in case: $X = 1, Y = 0$, with M unobserved.

Data types $\rightarrow$	X1M0Y0	X1M1Y0	Priors on causal types	Posteriors on causal types
Causal types $\downarrow$				
$\theta_1^X, \theta_{00}^M, \theta_{10}^Y$	1	0	0.02	0.0769
$\theta_1^X, \theta_{10}^M, \theta_{10}^Y$	1	0	0.02	0.0769
$\theta_1^X, \theta_{01}^M, \theta_{01}^Y$	0	1	0.08	0.3077
$\theta_1^X, \theta_{11}^M, \theta_{01}^Y$	0	1	0.04	0.1538
$\theta_1^X, \theta_{00}^M, \theta_{11}^Y$	0	1	0.02	0.0769
$\theta_1^X, \theta_{10}^M, \theta_{11}^Y$	0	1	0.02	0.0769
$\theta_1^X, \theta_{01}^M, \theta_{11}^Y$	1	0	0.04	0.1538

Data types →	X1M0Y0	X0M1Y0	Priors on causal types	Posteriors on causal types
$\theta_1^X, \theta_{11}^M, \theta_{11}^Y$	1	0	0.02	0.0769

For more complex models and estimands, it can be more difficult to eyeball the corresponding causal types. In practice, therefore, we use the `get_query_types` function in the `CausalQueries` package to do this for us.

```
X0.M10.Y10, X1.M10.Y10, X0.M01.Y10, X1.M01.Y10, X0.M10.Y01, X1.M10.Y01, X0.M01.Y01, X
```

This completes the abstract representation of the process tracing procedure. We now build up the intuition by walking through the procedure for simple mediation and moderation models.

6.1.3 Illustration with code

```
XMY <- make_model("X -> M -> Y") %>%
       set_parameters (c(.5, .5, .2, .2, .4, .2, .2, .2, .4, .2))

query_model(model = XMY,
              queries = list(PC = "Y[X=1] > Y[X=0]"),
              given = list(TRUE, "X==1 & Y==1", "X==1 & Y==1 & M==0", "X==1 & Y==1 &
              using = "parameters")
```

Query	Given	Using	mean
PC	-	parameters	0.200
PC	X==1 & Y==1	parameters	0.385
PC	X==1 & Y==1 & M==0	parameters	0.250
PC	X==1 & Y==1 & M==1	parameters	0.444

6.2 Five principles

6.2.1 Classic qualitative tests are special cases of updating on a model

The approach we have described here updates on the model given data on all variables, and from the model makes inferences to estimands.

This procedure appears different to the approach described, for example, in **?** and in Chapter 5, in which one seeks specific evidence that is directly informative about causal propositions: "clues" that are arise with different probabilities if one proposition or another is true. In fact however the approaches are deeply connected. This "probative value of clues" approach can indeed be *justified* by reference to more fully elaborated models of the world.

To see this we can write down the probability of observing $K = 1$ conditional on causal type X, using the ϕ notation from **?** and introduced in Chapter 5. Here ϕ_{jx} refers to the probability of observing a clue in a case of type j when $X = x$. Starting with our prior distribution over the lower-level causal types (the λ's), we can derive, for an $X = 1$ case, the probability of seeing the clue if the case is of type b (positive effect) or of type d (no effect, Y always 1):

$$\begin{aligned}\phi_{b1} &= \frac{\lambda^K_{01}\lambda^Y_{01}}{\lambda^K_{01}\lambda^Y_{01} + \lambda^K_{10}\lambda^Y_{10}} \\ \phi_{d1} &= \frac{\lambda^Y_{11}(\lambda^K_{01} + \lambda^K_{11}) + \lambda^K_{11}\lambda^Y_{01}}{\lambda^Y_{11} + \lambda^K_{00}\lambda^Y_{10} + \lambda^K_{11}\lambda^Y_{01}}\end{aligned} \tag{6.1}$$

These quantities allow for easy mapping between our prior beliefs about our causal query—as expressed in the lower level model—and the classic process-tracing tests in **?**. Figure **??** illustrates. In each panel, we manipulate a prior for one or more of the lower-level causal effects, keeping all other priors flat, and we see how probative value changes. As the curves for ϕ_b and ϕ_d diverge, probative value is increasing since there is an increasing difference between the probability of seeing the clue if X has a positive effect on Y and the probability of seeing the clue if X has no effect.

In the left panel, we see that as we place a lower prior probability on K's being negatively affected by X,[10] seeking $K = 1$ increasingly takes on the quality of a hoop test for X's having a positive effect on Y. The clue, that is, increasingly becomes something we must see if X positively affects Y, with the clue remaining moderately probable if there is no effect. Why? The less likely we believe it is that $K = 0$ was caused by $X = 1$, the less consistent the observation of $K = 0$ is with X having a positive causal effect on Y via

[10]For a given value of λ^K_{01}, we hold the other λ^K values equal by assigning a value of $(1 - \lambda^K_{01})/3$ to each.

K (since, to have such an effect, if $X = 1$ and $K = 0$, would precisely have to mean that $X = 1$ *caused* $K = 0$).

In the second graph, we simultaneously change the prior probabilities of zero effects at both stages in the sequence: of K and Y being 1 regardless of the values of X and K, respectively.[11] We see here that, as the probabilities of zero effects jointly diminish, seeking $K = 1$ increasingly becomes a smoking-gun test for a positive effect of X on Y: the probability of seeing the clue if the case is a d type diminishes. The reason is that, as zero effects at the lower level become less likely, it becomes increasingly unlikely that $K = 1$ could have occurred without a positive effect of X on K, and that $Y = 1$ could have occurred (given that we have seen $K = 1$) without a posiitve effect of K on Y.

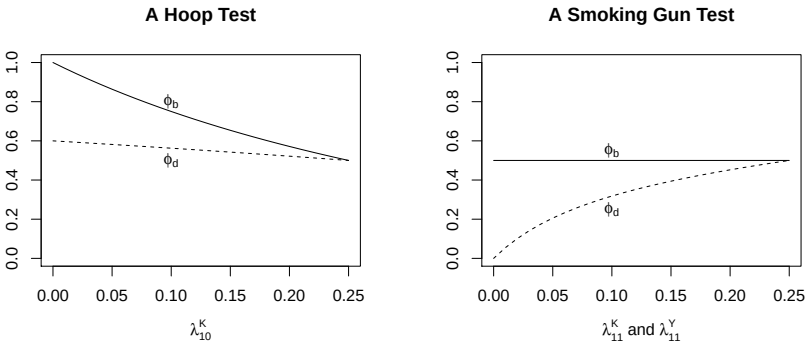

Figure 6.2: The probability of observing K given causal type for different beliefs on lower-level causal effects. In the left figure, priors on all lower-level causal effects are flat except for the probability that X has a negative effect on K. If we believe that it is unlikely that X has a negative effect on K, K becomes a 'hoop' test for the proposition that a case is of type b. The righthand figure considers simultaneous changes in λ^K_{11} and λ^Y_{11}—the probabilities that $K = 1$ regardless of X, and that $Y = 1$ regardless of K, with flat distributions on all other lower-level effects. With λ^K_{11}, λ^Y_{11} both close to 0, K becomes a 'smoking gun' test for the proposition that X has a positive effect on Y (b type).

[11] For a given value of λ^K_{11}, we hold the other λ^K's equal by assigning a value of $(1 - \lambda^K_{11})/3$ to each; likewise for λ^Y_{11} and the other λ^Y values.

6.2.2 A DAG alone does not get you probative value

Suppose that we start with the belief that any effect of X on Y must run through M, and that there is no confounding at any stage. This implies the simple $X \rightarrow M \rightarrow Y$ model. Suppose, further, that we have no prior knowledge about the distribution of nodal types and thus posit flat priors over θ^M and θ^Y. Now we would like to conduct process tracing and observe M to tell us about the effect of X on Y. In an $X = Y = 1$ case, for instance, is this model sufficient to allow the observation of M to provide leverage on whether $X = 1$ caused $Y = 1$?

It is not. We can learn nothing from about X's effect on Y from M. Observing a process is *uninformative* if all that we know is the structure of relations of conditional independence.

To see why at an inuitive level, consider that there are two causal types that will satisfy the query, $X = 1$ caused $Y = 1$. Those are the types $\theta^X_1\theta^M_{01}\theta^Y_{01}$ and $\theta^X_1\theta^M_{10}\theta^Y_{10}$: either linked positive effects or linked negative effects could generate an overall positive effect of X on Y. Moreover, with flat priors over nodal types, these causal types are equally likely. Now, think about what we would conclude if we collected process data and observed $M = 1$ in the $X = Y = 1$ case. This observation would rule out one way in which the query could be satisfied: the causal type with linked negative effects. And what if we observed, instead, $M = 0$? This would rule out the other way in which the query could be satisfued: linked positive effects. But since each of these causal types started out with equal weights in our priors, there can be no updating from eliminating one or the other.

6.2.3 Uncertainty does not alter inference for single case causal inference

In the procedure described for process tracing in this chapter (and different to what we introduce in Chapter 8) we assume that λ is known and we do not place uncertainty around it.

This might appear somewhat heroic, but in fact for single case inference it is without loss of generality. The expected inferences we would make for any estimand accounting for priors is the same as the inferences we if we use the expectation only.

To see this, let π_j denote the probability of observing causal type j and $p(D)$ te probability of observing data realization D. Say that $j \in D$ if type j produces data type D and say $j \in E$ if casual type j is an element of the estimand set of interest. For instance in an $X \rightarrow Y$ model, if we observe $X = Y = 1$ then D consists of causal types $D = (\theta_1^X, \theta_{01}^Y), (\theta_1^X, \theta_{11}^Y))$ and the estimand set for "X has a positive effect on Y" consists of $E = (\theta_1^X, \theta_{01}^Y), (\theta_0^X, \theta_{01}^Y))$.

The posterior on an estimand E given data D given prior over π, $p(\pi)$ is:

$$\Pr(E|D) = \int_\pi \frac{\sum_{j \in E \cap D} \pi_j}{\sum_{j \in D} \pi_j} f(\pi) d\pi$$

However, since for any π, $\sum_{j \in D} \pi_j = p(D)$ we have:

$$\Pr(E|D) = \int_\pi \sum_{j \in E \cap D} \pi_j f(\pi) d\pi / p(D) = \sum_{j \in E \cap D} \overline{\pi}_j / p(D)$$

6.2.4 Probative value requires $d-$connection

As we have argued, causal estimands can be expressed as the values of exogenous nodes in a causal graph. Case-level causal effects and causal paths can be defined in terms of response-type nodes; average effects and notable causes in terms of population-level parameter nodes (e.g., π or λ terms); and questions about actual causes in terms of exogenous conditions that yield particular endogenous values (conditioning on which makes some variable a counterfactual cause).

We thus define causal inference more generally as *the assessment of the value of one or more unobserved (possibly unobservable) exogenous nodes on a causal graph, given observable data.* To think through the steps in this process, it is useful to distinguish among three different features of the world, as represented in our causal model: there are the things we want to learn about; the things we have already observed; and the things we could observe. As notation going forward, let:

- $\mathcal{Q}$ denote the exogenous variables that define our *query*; we generally assume that $\mathcal{Q}$ cannot be directly observed so that its values must be