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.

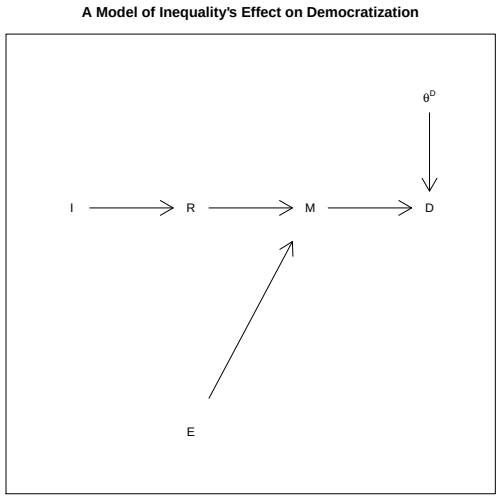

Figure 2.1: A simple causal model in which high inequality (I) affects the democratization (D) via redistributive demands and mass mobilization (M), which is also a function of ethnic homogeneity (E). The arrows show relations of causal dependence between variables. The graph does not capture the ranges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