Integrated Inferences

Macartan Humphreys and Alan M. Jacobs

Version: 19 November 2021

Contents

Preface

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.
- In Part IV we put models into question and outline a range of strategies one can use to justify and evaluate causal models.

We have developed an `R` package—`CausalQueries`—to accompany 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 models used in this book.

Chapter 1

Introduction

We describe the book's general approach, preview our argument for the utility of causal models as a framework for choosing research strategies and drawing causal inferences, and provide a roadmap for the rest of the book.

Here is the key idea of this book.

Quantitative social scientists spend a lot of time trying to understand causal relations between variables by looking across large numbers of cases to see how outcomes differ when potential causes differ. This strategy relies on variation in causal conditions across units of analysis, and the quality of the resulting inferences depends in large part on what forces give rise to that variation.

Qualitative social scientists, like historians, spend a lot of time looking at a smaller set of cases and seek to learn about causal relations by examining evidence of causal processes in operation within these cases. Qualitative scholars rely on theories of how things work, theories that specify what should be observable within a case if indeed an outcome were generated by a particular cause.

These two approaches seem to differ in what they seek to explain—individual-level or population-level outcomes; in the forms of evidence they require—cross-case variation or within-case detail; and in what they need to assume—knowledge of assignment processes or knowledge of causal processes.

The central theme of this book is that this distinction, though culturally real (**?**), is neither epistemologically deep nor analytically helpful. Social scientists can work with causal models that simultaneously exploit cross-case variation and within-case detail, that address both case-level and population-level questions, and that both depend on, and contribute to developing, theories of how things work.[1]

We describe an approach to doing this in which researchers *form* causal models, *update* those models using data, and then *query* the models to get answers to particular causal questions. This framework is very different from standard statistical approaches in which researchers focus on selecting the best estimator to estimate a particular estimand of interest. In a causal models framework, the model itself gets updated, not the estimate: we begin by learning about processes, and only then draw inferences about particular causal relations of interest, either at the case level or at the population level.

We do not claim that a causal-model-based approach is the best or only strategy suited to addressing causal questions. There are plenty of settings in which other approaches would likely work better. But we do think that the approach holds considerable promise—allowing researchers to combine disparate data in a principled way to ask a vast range of causal questions, helping integrate theory and empirics in a compelling way, and providing coherent guidance on research design—and that it should have a place in the applied researcher's toolkit.

Our goals in this book are to motivate this approach; provide an introduction to the theory of structural causal models; provide practical guidance for setting up, updating, and querying causal models; and show how the approach can inform key research-design choices, especially case-selection and data-collection strategies.

1.1 The Case for Causal Models

There are three closely related motivations for embracing a causal models approach. One is is a concern over the limits of design-based inference. A

[1] Indeed, though we will sometimes follow convention and refer to "within case" and "cross-case" observations, all data are data on cases and enter into analysis in the same fundamental way: we are always asking how consistent a given data pattern is with alternative sets of beliefs.

second is an interest in integrating qualitative knowledge with quantitative approaches. A third is an interest in better connecting empirical strategies to theory.

1.1.1 The limits to design-based inference

To caricature positions a bit, consider the difference between an engineer and a skeptic. The engineer tackles problems of causal inference using models: theories of how the world works, generated from past experiences and applied to the situation at hand. They come with prior beliefs about a set of mechanisms operating in the world and, in a given situation, will ask whether the conditions are in place for a known mechanism to operate effectively. The skeptic, on the other hand, maintains a critical position, resisting the importation of beliefs that are not supported by evidence in the case at hand.

The engineer's approach echoes what was until recently a dominant orientation among social scientists. At the turn of the current century, much analysis—both empirical and theoretical—took the form of modelling processes ("data generating processes") and then interrogating those models.

Over the last two decades, however, skeptics 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 which provide a clear language for thinking about causation (see **?**, **?** among others) without having to invoke fully specfied models of data-generating processes.

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

Excitement about the strengths of these approaches has been mixed with various concerns regarding how the approach shapes inquiry. We highlight two.

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 both 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 which, or situations in which, exogeneity can be established "by design" would represent a dramatic narrowing of social science's ken. To be clear, this is not an argument against experimentation or design-based inference when these can be used; rather 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. Yet often the model is the thing we want to learn about. Our goal as social scientists is to come to grips with how the world works, not simply to collect propositions about the effects that different causes have had on different outcomes 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 out of the equation, as it were, we 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. On our read of this literature, the logic of process tracing in these accounts depends on a particular form of model-based inference.[2] While

[2] 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

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 in political science and sociology 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 observation of covariation of causes and effects across cases.

The process-tracing literature has also advanced increasingly 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 tests", "smoking gun tests", etc.) involving varying degrees of specificity and sensitivity (**?**, **?**).[3] 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 (**?**, **?**, **?**).[4]

Yet, conceptualizing the different ways in which probative value might operate leaves a fundamental question unanswered: what gives within-case

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.

[3] 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 **?**).

[4] 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.

evidence its probative value with respect to causal relations? We do not see a clear answer to this question in the current process-tracing literature. Implicitly—but worth rendering explicit—*the probative value of process-tracing evidence depends on researcher beliefs that come from 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 the role that models can—and, in our view, must—play in drawing case-level causal inferences and to clarify conditions under which these models can be defended. To do so we draw on an approach to specifying causal models developed originally in computer science and that predates this work in qualitative methodology. The broad approach, described in **?** and **?**, is consistent with the potential outcomes framework, and provides rules for updating on population and case level causal queries from different types of data.

In addition to clarifying the logic of qualitative inference, we will argue that such causal models can also enable the systematic integration of qualitative and quantitative forms of evidence. Social scientists are increasingly developing mixed-method research designs, 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 cases. Now-classic examples of this approach include 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 (**?**; **?**). 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 (**?**).

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 has not supplied specific principles for aggregating findings—whether mutually reinforcing or contradictory—across different modes of analysis.[5] As we aim to demonstrate in this book, however, grounding inference in causal models provides a very natural way of combining information of the X, Y variety with information about the causal processes connecting X and Y. The approach that we develop here can be readily addressed both to the case-oriented questions that tend to be of interest to qualitative scholars and to the population-oriented questions that tend to motivate quantitative inquiry.

As will become clear, 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 because all causal inference depends fundamentally on covariation but because in a causal-model-based inference, what matters for the informativeness of a piece of evidence is how that evidence alters beliefs about a model, and in turn, a query. 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

The relationship between theory and empirics has been a surprisingly uncomfortable one in political science. In a recent intervention, for instance, **?** draw attention to and critique political scientists' widespread reliance on the "hypothetico-deductive" (H-D) framework, in which a theory or model is

[5] 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).

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 H-D framework, including with the idea that the truth of a model logically derived from first principles can be *tested* against evidence.

In fact the relationship between theory and evidence in social inquiry is often surprisingly unclear both in 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 is not clear, however, is a 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? 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 in a case, estimating an average causal effect, or identifying the pathway through which an effect is generated? Perhaps most confusingly, if the theory tells us what to look for to draw an inference, can the inferences be about the theory itself or are we constrained to make theory dependent inferences? In short, how exactly can we ground causal inferences from within-case evidence in background knowledge about how the world works?

Much 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 less regard for intervening or surrounding causal relationships.[6] If a theory suggests a *set* of relations, it is common to examine these separately—does A cause B does B cause C? are relations stronger or weaker here or there?—without standard procedures for bringing the disparate pieces of evidence together to form theoretical conclusions. More attention has been paid to empirical implications of theoretical models than to theoretical implications of empirical models.

In this book, we seek to show how scholars can simultaneously make fuller and more explicit use of theoretical knowledge in designing their research projects and analyzing data and make use of data to update on theoretical models. Like Clarke and Primo, we treat models not as veridical accounts of the world but as maps: 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 (though, in later chapters, we do discuss ways of justifying and evaluating models). Rather, our focus is on how we can systematically *use* causal models—in the sense of *mobilizing background knowledge of the world*—to guide our empirical strategies and inform our inferences. Grounding our empirical strategy in a model allows us, in turn, to learn about features of the model itself as we encounter the data.

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 (see especially **?**). 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 with which qualitative and quantitative scholars routinely grapple.[7]

[6] There are of course many exceptions, including work that uses structural equation modeling, and research that focuses specifically on understanding heterogeneity and mediation processes.

[7] 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

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 benefits of this approach are multiple. In particular, the framework that we develop in this book provides:

- A grounding for assessing the "probative value" of evidence drawn from different parts of any causal network. The approach yields a principled and transparent approach to answering the question: how should the observation of a given piece of data affect my causal beliefs about a case?

- A transparent, replicable method of aggregating inferences from observations drawn from different locations in a causal network. Having collected multiple pieces of evidence from different parts of a causal process or case context, what should I end up believing about the causal question of interest?

- A common approach for assessing a wide variety of queries (estimands). We can use the same apparatus to learn *simultaneously* about different case-level causal questions, such as "What caused the outcome in this case?" and "Through what pathway did this cause exert its effect?"

- Guidance for research design. Given finite resources, researchers must make choices about where to look for evidence. A causal model framework can help researchers assess, a priori, the relative expected informativenss of different evidentiary and case-selection strategies, conditional on how they think the world works and the question they want

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.

to answer.

The approach also offers a range of distinctive benefits to researchers seeking to engage in mixed-method inference and to learn about general causal relations, as well as about individual cases. The framework's central payoff for multi-method research is the systematic integration of qualitative and quantitative information to answer any given causal query. We note that the form of integration that we pursue here differs from that offered in other accounts of multi-method research. In **?**'s approach, for instance, one form of data—quantitative *or* qualitative—is always used to draw causal inferences, while the other form of data is used to test assumptions or improve measures employed in that primary inferential strategy. In the approach that we develop in this book, in contrast, we are always using *all* information available to update on causal quantities of interest. In fact, within the causal models framework, there is no fundamental difference between quantitative and qualitative data, as both enter as values of nodes in a causal graph. This formalization—this reductive move—may well discomfit some readers. And we acknowledge that our approach undeniably involves a loss of some of what makes qualitative research distinct and valuable. Yet, this translation of qualitative and quantitative observations into a common, causal model framework offers major advantages. Beyond the integration of different forms of information, these advantages include:

- Transparency. The framework makes manifest precisely how each form of evidence enters into the analysis and shapes conclusions.

- Learning across levels of analysis. In a causal model approach, we use case-level information to learn about populations and general theory. At the same time, we use what we have learned about populations to sharpen our inferences about causal relations within individual cases.

- Cumulation of knowledge. A causal model framework provides a straightforward, principled mechanism for building on what we have already learned. As we see data, we update our model; and then our updated model can inform the inferences we draw from the next set of observations. Models can, likewise, provide an explicit framework for positing and learning about the generalizability and portability of findings across research contexts.

- Guidance for research design. With a causal model in hand, we can formally assess key multi-method design choices, including the balance we should strike between breadth (the number of cases) and depth (intensiveness of analysis in individual cases) and the choice of cases for intensive analysis.

Using causal models also has substantial implications for common methodological intuitions, advice, and practice. To touch on just a few of these implications:

- 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 about "on the regression line" cases or those in which the outcome occurred, and there is nothing necessarily damning about selecting on the dependent variables. Rather optimal case selection depends on factors that have to date received little attention, such as the population distribution of cases and the probative value of the available evidence.
- Our 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 here is not new. Rather, we see the book's goals as being of three kinds. First, we aim to import insights: 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 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, we draw 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, we provide a set of usable tools for implementing the approach. We provide intuition and software, the `CausalQueries` package, that researchers can use to make research design choices and draw inferences from the data.

There are also important limits to this book's contributions and aims. First, while we make use of Bayesian inference throughout, we do not engage here with fundamental debates over or critiques of Bayesianism itself. (For excellent treatments of some of the deeper issues and debates, see, for instance, **?** and **?**.)

Second, this book does not address matters of data-collection (e.g., conducting interviews, searching for archival documents) or the construction of measures. For the most part, we assume that data are either at hand or can be gathered, and we bracket the measurement process itself. That said, a core concern of the book is using causal models to identify the *kinds* of evidence that qualitative researchers will want to collect. In Chapter **??**, we show how causal models can tell us whether observing an element of a causal process is potentially informative about a causal question; and in Chapter **??** we demonstrate how we can use models to assess the likely learning that will arise from different clue-selection strategies. We also address the problem of measurement error in Chapter **??**, showing an approach to using causal models to learn about error from the data.

Finally, while we will often refer to the use of causal models for "qualitative" analysis, we do not seek to assimilate all forms of qualitative inquiry into a causal models framework. Our focus is on work that is squarely addressed to matters of causation; in particular, the logic that we elaborate is most closely connected to the method of process tracing. More generally, the formalization the we introduce here—the graphical representation of beliefs and the application of mathematical operations to numerically coded observations—will surely strike some readers as reductive and not particularly "qualitative." It is almost certainly the case that, as we formalize, we leave behind some of what makes qualitative research distinctive and valuable. Our aim in this book is not to discount the importance of those aspects of qualitative inquiry that resist formalization, but to show some of things we *can* do if we are willing to formalize.

1.3 The Road Ahead

The book is divided into four main parts.

In the first part of the book, we set out the basics. In Chapter **??**, following a review of the common potential-outcomes approach to causality, we introduce the concept and key components of a causal model. Chapter **??** illustrates how we can represent of causal beliefs in the form causal models by translating the arguments of a several prominent works of political science into causal models. In Chapter **??**, we set out a range of causal questions that researchers might want to address—including questions about case-level causal effects, population-level effects, and mechanisms—and define these queries within a causal model framework. Chapter **??** offers a primer on the key ideas in Bayesian inference that we will mobilize in later sections of the book. In Chapter **??**, we map between causal models and theories, showing how we can think of any causal model as situated within a hierarchy of complexity: within this hierarchy, any causal model can be justified by references to a "lower level", more detailed model that offers a theory of why things work the way do that the higher level. This conceptualization is crucial insofar as we use more detailed (lower-level) models to generate empirical leverage on relationships represented in simpler, higher-level models.

The second part of the book shows how we can use causal models to undertake process-tracing and mixed method inference. Chapter **??** lays out the logic of case-level inference from causal models: the central idea here is that what we learn from evidence is always conditional on the prior beliefs embedded in our model. In Chapter **??**, we illustrate model-based process-tracing with an application to the substantive issue of economic inequality's effects on democratization. Chapter **??** moves to mixed data problems: situations in which a researcher wants to use "quantitative" (broadly, X, Y) data on a large set of cases and more detailed ("qualitative") data on some subset of these cases. We show how we can use any arbitrary mix of observations across a sample of any size (greater than 1) to update on all causal parameters in a model, and then use the updated model to address the full range of general and case-level queries of interest. In Chapter **??**, we illustrate this integrative approach by revisiting the problem of inequality and democracy introduced in Chapter **??**. Finally, in Chapter **??**, we take the project of integration a step further by showing how we can use models to integrate findings across *studies* and across *settings*. We show, for instance, how we can learn jointly from the data generated by an observational study and an experimental study of the same causal domain and how models can help us reason in principled ways about the transportability of findings across contexts.

The third part of the book unpacks what causal models can contribute to research design. Across Chapters **??**, **??**, and **??** we demonstrate how researchers can mobilize their models, as well as prior observations, to determine what kind of new evidence is likely to be most informative about the query of interest, how to strike the balance between extensiveness and intensiveness of analysis, and which cases to select for in-depth process tracing.

The fourth and final part of the book steps back to put the model-based approach into question. Until this point, 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 advocate a focus on models, we know that skeptics are right to distrust them. This final part approaches this problem from two perspectives. In Chapter **??**, we demonstrate the *possibility* of justifying models from external evidence, though we do not pretend that the conditions for doing so will arise commonly. In Chapter **??**, drawing on common practice in Bayesian statistics, we present a set of strategies that researchers can use to evaluate and compare the validity of models, and to investigate the degree to which findings hinge on model assumptions.

In the concluding chapter we summarize what we see as the main advantages of a causal-model-based approach to inference, draw out a set of key concerns and limitations of the framework, and identify what we see as the key avenues for progress in model-based inference.

Here we go.

Part I

Foundations

Chapter 2

Causal Models

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

Causal claims are everywhere. Causal knowledge is often the end goal of empirical social science. It is also a key *input* into causal inference.[1] Causal assumptions are also hidden in seemingly descriptive statements: claims that someone is guilty, or exploited, or powerful, or weak, involve beliefs about how things would be were conditions different. Even when scholars carefully try to avoid causal claim-making, causal verbs—depends, drives, produces, influences—tend to surface.

But while causal claims are commonplace, it is not always clear (1) what exactly is meant by a causal relation and (2) how causal knowledge about one thing can be marshaled to justify causal claims about another. For our purposes, the counterfactual view of causality addresses the first question. Causal models address the second.

2.1 The counterfactual model

We begin with what we might think of as a meta-model, the counterfactual model of causation. At its core, a counterfactual understanding of causation

[1] As nicely put by **?**: no causes in, no causes out. We return to the point more formally later.

captures a simple notion of causation as "difference-making."[2] In the counterfactual view, to say that X caused Y is to say: *had* X been different, Y *would have been* different.

The causal effect, in this view, is the difference between two things that might have happened. This means that *by definition, causal effects are not measurable quantities*. They are not differences between possible observations in the world, but, at best, differences between outcomes in the world and counterfactual outcomes. They need to be inferred not measured.

Moreover, in this view, 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 somehow have *made* X different, Y would have been different. In the terminology of **?**, we represent this quantity using a "do" operator: $Y(do(X = x))$ is the value of Y when x is *set* to x.

Consider a simple example. Students with teacher A perform well without studying. Students with teacher B perform well if they study, and do not perform well if they do not study. Moreover, only students with teacher B in fact study. And all perform well.

When we say that one of teacher B's students did well *because* they studied, we are comparing the outcome that they experienced to the outcome they would have experienced if they had had teacher B (as they did) but (counterfactually) had not studied. Notably, we are *not* comparing their realized outcome to the outcome they would have experienced if they had been among the people that in fact didn't study (i.e., if they had had teacher A).

A second example. Consider the claim that Switzerland democratized ($D = 1$) because it had a relatively low level of economic inequality ($I = 0$) (drawing on the logic of **?**). In the counterfactual view, this is equivalent to saying that, had Switzerland *not* had a high level of equality, the country would not have democratized. High economic equality made a difference. The

[2] The approach is sometimes attributed to David Hume, whose writing contains ideas both about causality as regularity and causality as counterfactual. On the latter, Hume's key formulation 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 **?**.

comparison for the causal statement is with the outcome Switzerland would have experienced under an intervention that boosted its historical level of economic inequality—*not* with how Switzerland would have performed if it had been one of the countries that *in fact* had higher levels of inequality, cases that likely differ from Switzerland in other causally relevant ways.

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, and 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. We need a language to describe these different types of relations.

2.1.1 Potential outcomes

The "potential outcomes" framework is useful for describing 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.

A setting in which it is quite natural to think about potential outcomes is medical treatment. Imagine some individuals in a diseased 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 ($Y = 0$). Given the assignments of all other individuals,[3] we can treat each individual as belonging to one of four unobserved response "types," defined by the outcomes that the individual *would have* if they received or did not receive treatment:[4]

[3] We noted that we are conditioning on the assignments of others. If we wanted to describe outcomes as a function of the *profile* of treatments received by others we would have a more complex types space. For instance in an $X \rightarrow Y$ model with 2 individuals we would report how (Y_1, Y_2) respond to (X_1, X_0); each vector can take on four values producing a type space with 4^4 types rather than 2^2. The complex type space could be reduced back down to four types again, however, if we invoked the assumption that the treatment or non-treatment of one patient has no effect on the outcomes of other patients—an assumption known as the stable unit treatment value assumption (SUTVA).

[4] See **?**} for an early classification of this form. The literature on probabilistic models also refers to such strata as "canonical partitions" or "equivalence classes."

- **a**dverse: Those individuals 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

Table **??** maps the four types (a, b, c, d) onto their respective potential outcomes. 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. In each cases we are imagining *controlled* changes in treatment: the responses if treatments are changed without changes to other background conditions about the case.

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	**b**eneficial	**c**hronic	**d**estined
Outcome if not treated	Healthy	Sick	Sick	Healthy
Outcome if treated	Sick	Healthy	Sick	Healthy

We highlight that, in this framework, case-level causal relations are treated as deterministic. A given case has a set of potential outcomes. Any uncertainty about outcomes enters as incomplete knowledge of a case's "type," not from underlying randomness in causal relations. 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 also 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. There are certainly situations we could imagine in which one might want to conceptualize potential outcomes themselves as random (for instance, if individuals in different conditions play different lotteries). But for the vast majority of the settings we condsider, not much of importance is lost if we treat potential outcomes as deterministic but possibly unknown: at the end of the day something will occur or it will not occur, we just do not know which it is.

2.1.2 Generalization

Throughout the book, we generalize from this simple setup. 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), there are only four sets of possible potential outcomes, or "types." More generally, for variable, Y, we will use θ^Y to capture the unit's "type": the way that Y responds to its potential causes.[5] We, further, add subscripts to denote particular types. Where there are four possible types, for instance, we use the notation θ^Y_{ij}, where i represents the case's potential outcome when $X = 0$ and j is the case's potential outcome when $X = 1$.

Adopting this notation, for a causal structure with one binary causal variable and a binary outcome, the four types can be represented as $\{\theta^Y_{10}, \theta^Y_{01}, \theta^Y_{00}, \theta^Y_{11}\}$, as shown in Table **??**:

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 the matter of inequality and democratization to illustrate, let $I = 1$ represent a high level of economic equality and $I = 0$ its absence; let

[5] Later, we will refer to these as "nodal types."

$D = 1$ represent democratization and $D = 0$ its absence. A θ^D_{10} (or a) type is a 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 a case, i, of θ^D_{10} type is $\tau_i = -1$. A θ^D_{01} type (or 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 $\tau_i = 1$). A θ^D_{00} type (c type) is a case that will not democratize regardless of the level of equality (effect $\tau_i = 0$); and a θ^D_{11} type (d type) is one that will democratize regardless of the level of equality (again, effect $\tau_i = 0$).

In this setting, a causal *explanation* of a given case outcome amounts to a statement about its type. The claim that Switzerland's high level of equality was a cause of its democratization is equivalent to saying that Switzerland democratized and is a θ^D_{01} type. To claim that Sierra Leone democratized because of low inequality is equivalent to saying that Sierra Leone democratized and is a 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).

Now, let us consider more complex causal relations. Suppose now that 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 with the ordering (and, in general, mapping any two "parents" alphabetically into X_1 and X_2). The next equation connects this notation to the "do" notation used to convey the idea that conditions are controlled.

$$Y_{hijk} \left\{ \begin{array}{lcl} h & = & Y|do(X_1 = 0, X_2 = 0) \\ i & = & Y|do(X_1 = 1, X_2 = 0) \\ j & = & Y|do(X_1 = 0, X_2 = 1) \\ k & = & Y|do(X_1 = 1, X_2 = 1) \end{array} \right. \tag{2.1}$$

Thus, for instance, θ^Y_{0100} means that Y is 1 if $X_1 = 1$ and $X_2 = 0$ and is 0

Table 2.3: With two binary causal variables, there are 16 causal types: 16 ways in which Y might respond to changes in the other 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 $X_1=1
θ^Y_{0000}	0	0	0	0
θ^Y_{1000}	1	0	0	0
θ^Y_{0100}	0	1	0	0
θ^Y_{1100}	1	1	0	0
θ^Y_{0010}	0	0	1	0
θ^Y_{1010}	1	0	1	0
θ^Y_{0110}	0	1	1	0
θ^Y_{1110}	1	1	1	0
θ^Y_{0001}	0	0	0	1
θ^Y_{1001}	1	0	0	1
θ^Y_{0101}	0	1	0	1
θ^Y_{1101}	1	1	0	1
θ^Y_{0011}	0	0	1	1
θ^Y_{1011}	1	0	1	1
θ^Y_{0111}	0	1	1	1
θ^Y_{1111}	1	1	1	1

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 we read types off of four-digit subscripts. For instance, the table shows us that for nodal type θ^Y_{0101}, X_1 has a positive causal effect on Y but X_2 has no effect. On the other hand, for type θ^Y_{0011}, X_2 has a positive effect while X_1 has none.

We also capture interactions here. For instance, in a θ^Y_{0001} type, X_2 has a positive causal effect if and only if X_1 is 1. In that type, X_1 and X_2 serve as "complements." For θ^Y_{0111}, X_2 has a positive causal effect if and only if X_1 is 0. In that setup, X_1 and X_2 are "substitutes."

This is a rich framework in that it allows for all possible ways in which a set of multiple causes can interact with each other. Often, when seeking to explain the outcome in a case, researchers proceed as though causes are necessarily *rival*, 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 can be 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. This typological, potential-outcomes conceptualization provides a straightforward way of representing this kind of complex causation.

Because of this complexity, when we say that X caused Y in a given case, we will generally mean that X was *a* cause, not *the* (only) cause. 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.

We will mostly use θ^Y_{ij}-style notation in this book to refer to types. We will, however, occasionally revert to the simpler a, b, c, d designations when that helps ease exposition. As types play a central role in the causal-model framework, we recommend getting comfortable with both forms of notation before going further.

Using the same framework, we can generalize to structures in which a unit has any number of causes and also to cases in which causes and outcomes are non-binary. As one might imagine, the number of types increases rapidly (very rapidly) as the number of considered causal variables increases, as it also does as we allow 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 $2^{(2^n)}$ types of this form: that is, $k = 2^n$ combinations of values of causes to consider, and 2^k distinct ways to react to each combination. If causes and outcomes are ternary instead of binary, we have $3^{(3^n)}$ causal types of this form. Yet, the basic principle of representing possible causal relations as patterns of potential outcomes remains unchanged, at least as long as variables are discrete.

2.1.3 Summaries of potential outcomes

So far, we have focused on causal relations at the level of an individual case. Causal relations at the level of a population are, however, simply a summary of causal relations for cases, and the same basic ideas can be used. 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 quite 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. Such summaries allow us to move from discussion of the cause of a single outcome to discussions of average effects, a distinction that we take up again in Chapter **??**.

2.2 Causal Models and Directed Acyclic Graphs

So far we have discussed how a single outcome is affected by one or more possible causes. However, these same ideas can be used to describe more complex relations between collections of variables—for example, with one variable affecting another directly as well as indirectly via its impact on some mediating variable.

Potential outcomes tables can be used to describe such complex relations. 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., **?** and **?**), and we introduce here a set of basic principles that readers will need to keep in mind in order to follow the argumentation in this book.

As we shift to talking about networks of causal relations between variables we will also shift our language. When talking about causal networks, or causal graphs, we will generally refer to variables as "nodes." And we will sometimes use familial terms to describe relations between nodes. For instance, if A is a cause of B we will refer to A as a parent and B as a child and represent this with an arrow from A to B. If X_1 and X_2 have a child in common (both directly affecting the same variable) we refer to them as "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).

Return to our running democratization example, but suppose now that we have more fully specified beliefs about how the level of economic inequality can have an effect on whether a country democratizes. We might believe that 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 the capacities of lower-income citizens for collective action. In this model R is a parent of M, I is an ancestor of M but not a parent. R and E are spouses and M is their child. We can visualize this model as a Directed Acyclic Diagram (DAG) in Figure **??**.

Fundamentally, we treat causal models in this book as formal representations of *beliefs* about how the world works—or, more specifically, about causal relations within a given domain. We might use a causal model to capture our own beliefs, a working simplification of our beliefs, or a set of potential beliefs that one might hold. The formalization of *prior* beliefs in the form of a causal model is the starting point for research design and inference in this book's analytic framework. Using the democratization example, we will now walk through the three components of a causal model in which our beliefs get embedded: nodes, functions, and distributions.

2.2.1 Components of a Causal Model

The three components of a causal model are (i) the nodes—that is, the set of variables we are focused on and how are they defined (ii) the functional relations—which nodes are "explained" by which other nodes and how, and (iii) probability distributions over unexplained elements of a model.

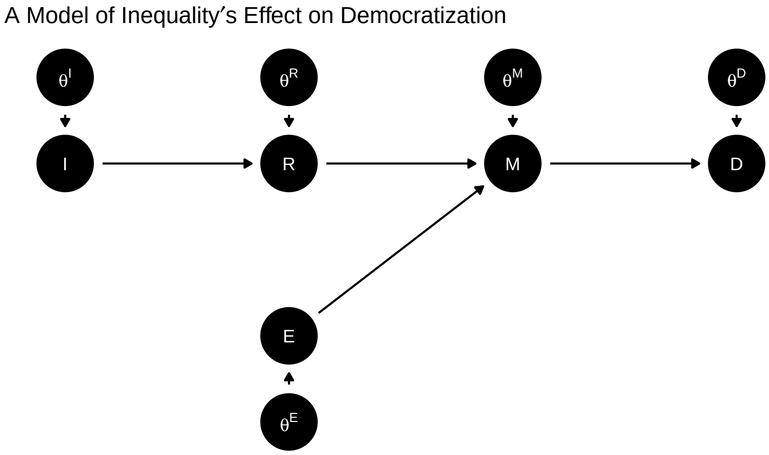


Figure 2.1: A simple causal model in which high inequality (I) affects the democratization (D) via redistributive demands and mass mobilization (M), which is also a function of ethnic homogeneity (E). The arrows show relations of causal dependence between variables. The graph does not capture the ranges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

2.2.1.1 The nodes

The first component of a causal model is the set of variables (nodes) across which the model characterizes causal relations. On the graph in Figure **??**, the five included variables are represented by the five lettered nodes. (In addition we mark θ^D on the graph though, as will be made clear, we will not think of this as a variable.)

In identifying the nodes, we also need to specify the *ranges* over which they can vary. We might specify, for instance, that all nodes in the model are binary, taking on the values 0 or 1. We could, alternatively, define a set of categories across which a node ranges or allow a node to take on any real number value or any value between a set of bounds. [6]

Notice that some of these nodes have arrows pointing *into* them: R, M, and D are endogenous nodes, meaning that their values are determined entirely

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

by other nodes in the model.

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

The θ terms require a little more explanation since they do not describe substantive nodes. In our discussion above, we introduced θ notation for representing types. Generally we can think of θ terms on a causal graph as unobservable and unspecified features of a causal domain that affect outcomes. These might include random processes (noise) or contextual features that we are unable to identify or do not understand. We imagine them pointing into everynode, whether indicated or not. In all cases the work they do is to characterize what the value of a node will be, given the value of its parents.

We note that our notation for representing these unobservable, unspecified influences differs from that commonly found in the literature on causal models. In many treatments, these components are themselves referred to as "exogenous" variables, and often labelled as sets $\mathcal{U}$, to be distinguished from the endogenous—named–variables often labelled as $\mathcal{V}$. We will generally use θ to denote these unobserved, unspecified influences to emphasize their particular role, as direct objects of interest in causal inquiry. As we will show, we can think of θ nodes as capturing the functional relations between endogeneous variables and as being quantities of direct interest for causal inquiry. We more fully develop this point—returning to the notion of θ terms as receptacles for causal effects—below.

2.2.1.2 The functions

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

The DAG already represents a critical part of these beliefs: the arrows, or directed edges, tell us *which nodes we believe to be direct causal inputs into other nodes.* 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. As we have said, 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) and ethnic heterogeneity (E) as shaped solely by factors exogenous to the model, captured by θ^I and θ^E, respectively.

Beyond the beliefs captured by the arrows in a DAG, we can express more specific beliefs about causal relations in the form of a causal function.[7] Specifying a function means writing down whatever general or theoretical knowledge we have about the direct causal relations between nodes. A function specifies how the value that one node takes on is determined by the values that other nodes—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. We take a non-parametric approach in this book, but it is helpful to juxtapose that approach to a parametric one.

Parametric approaches. A parametric approach specifies a functional form that relates parents to children. For instance, we might model one node as a linear function of another and write $D = \alpha + \beta M$, where β is a parameter that we do not know the value of at the outset of a study but about which we wish to learn. 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 might write: $D = \alpha + \beta M + \theta^D$. In this function, α, β might be the parameters of interest, with θ^D treated merely as a random disturbance. We can be still more agnostic by, for example, including parameters that govern how other parameters operate. Consider, for instance the function, $D = \beta M^{\theta^D}$. Here, D and M are linearly related if $\theta^D = 1$, but not otherwise.

Note 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 nodes are related.

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

The non-parametric approach. With discrete data, causal functions can take fully *non-parametric* form, allowing for *any possible relation* between parents and children. We use this framework for most of this book and thus spend some time developing the approach here.

We begin by returning to the concept of types. Drawing on our original four types from earlier in this chapter, we know that we can fully specify causal relations between a binary M and a binary D by allowing the node θ_D to range across four possible values $\{\theta^D_{10}, \theta^D_{01}, \theta^D_{00}, \theta^D_{11}\}$. For instance, θ^D_{10} represents a negative causal effect of M on D while θ^D_{00} represents D remaining at 0 regardless of M. Put differently, θ^D represents the non-parametric function that relates M to D. We can formally specify D's behavior as a function of M and θ^D in the following way:

$$D(M, \theta^D_{ij}) = \begin{cases} i & \text{if} \quad M = 0 \\ j & \text{if} \quad M = 1 \end{cases}$$

Note that θ^D ranges over *all possible* functional forms between these two binary variables.

How should we think about what kind of *thing* θ^D is, in a more substantive sense? It is probably most helpful to think of θ^D as an unknown and possibly random 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. Importantly, however, while we might think of θ^D as an unknown or random quantity, in this formulation θ^D should not be thought of as a nuisance—as "noise" that we would like to get rid of—but as *the quantity that we want to learn about*: we want to know whether M likely had a positive, negative, or no effect on D. We elaborate on this point at much greater length in Chapter **??**.

We can similarly use θ terms to capture causal relations involving any number of parent nodes. Every substantively defined node, J, in a graph can be thought of as having a θ^J term pointing into it, and the (unobservable) value of θ^J represents the mapping from J's parents (if it has any) to the value of J.

Applied to the binary nodes in Figure **??**, θ^J ranges as follows:

- **Nodes with no parents**: For an exogenous node, like I or E, θ^J represents an external "assignment" process can take on one of two values, θ^J_0, meaning that J is "assigned" to 0 or θ^J_1, meaning that J is assigned to 1. For instance, θ^I_0 typifies a case in which exogenous forces have generated low inequality.
- **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**: M has two parent nodes. Thus, θ^M will take on a possible 16 values of the form θ^M_{hijk} (θ^M_{0000}, θ^M_{0001}, etc.), using the syntax detailed earlier in this chapter.

Nodal types and causal types. For analytic applications later in the book, we will want to be able to think both about the type operating at a particular *node* and about *collections* of types operating across a model. We thus refer to θ^J as a unit's *nodal* causal type, or simply nodal type, for J. 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 the vector θ.

If we hypothetically knew a unit's causal type—all nodal types for all nodes—then we would know everything there is to know about that unit. Since the nodal types of exogenous nodes include *values* of all exogenous nodes, and the nodal types of all endogenous nodes specify how those nodes respond to all of their parents, a unit's causal type fully specifies all nodal values as well as all *counterfactual* nodal values for a unit.

We will sometimes refer to the values of θ as a unit's *context*. This is because θ captures all exogenous forces acting on a unit. This includes the assignment process driving the model's exogenous nodes (in our example, θ^I and θ^E) as well as all contextual factors that shape causal relations between nodes (θ^R, θ^M, and θ^D). Put differently, θ captures both how a unit reacts to situations and which situations it is reacting to. One implication is that there is no *formal* distinction between a unit's type and a unit's situation—between, say, a hungry person, and a person who has had no food.

Nodal types, causal types

term	symbol	meaning
nodal type	θ^J	The way that node J responds to the values of its parents. Example: θ^Y_{10}: 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: $(\theta^X_0, \theta^Y_{00})$, is a causal type that has $X = 0$ and that has $Y = 0$ no matter what the value of X.

A few important aspects of causal functions are worth highlighting.

1. 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, *directional* statements, with causes on the righthand side and an outcome on the left.

2. The collection of simple functions that map from the values of parents of a given node to the values of that node are sufficient to represent potentially complex webs of causal relations. For each node, 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 nodes pointing directly into it. Our outcome of interest, D, may be 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 parent—its direct cause, M. We then, separately, express a belief about how M is shaped by *its* parents, R and E. A node's function must include as inputs all, and only, those nodes that point directly into that node.[8]

3. As in the general potential-outcomes framework, all relations in a causal model are conceptualized as deterministic at the case level. Yet, there is not as much at stake here as one might think at first; by this

[8] The set of a node's parents is required to be minimal in the sense that a node 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.

> we simply mean that a node's value is *determined* by the values of its parents *along with* any stochastic or unknown components. We express uncertainty about causal relations, however, as unknown parameters, such as the causal types θ.

2.2.1.3 The distributions

Putting causal structure and causal functions together gives us a *structural causal model.* In a structural causal model, all endogenous nodes are, either directly or by implication, functions of a case's context (the values of the set of exogenous nodes).[9] What we have not yet inscribed into the model, however, is beliefs about how *likely* or *common* different kinds of contexts might be.

Thus, for instance, a structural causal model consistent with Figure **??** stipulates which nodes may have effects on which other nodes. But it says nothing in itself about the distribution of values of either the exogenous nodes or of the causal relations between nodes.[10] We have not said anything, for instance, about how common high inequality is across the relevant domain of cases or how common ethnic homogeneity is. Put differently, we have said nothing about the *distribution* of θ^I or of θ^E. Similarly, we have said nothing yet about how commonly mobilization has positive, negative, or null effects of democratization—that is, the distribution of θ^D—or about how commonly I and E have different possible joint causal effects on M (the distribution of θ^M).

In many research situations, we will have or want to posit a set of prior beliefs about how the world works under different conditions and about what kinds of conditions are more likely than others. We can express these beliefs about context as probability distributions over the model's θ^J terms. For instance, our structural causal model might tell us that E and R can jointly affect M.

[9] 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 node 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.

[10] 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.

We might, then, add to this a belief about θ^M such that, in the population of interest, redistribution rarely has a positive effect on mobilization when ethnic homogeneity is low. We would thus be putting a low probability on the nodal types for M in which R has a positive effect on M when $E = 0$, relative to M's other nodal types.[11]

We might add to this the belief that $E = 1$ in only 10% of cases in the population of interest, thus setting a 0.1 probability on θ^E_1. Note that these two beliefs jointly imply that R will rarely have a positive effect on M.

As with functions, we can also (and typically would) build uncertainty into our beliefs about the shares of different nodal types in the population. We do this by specifying a *distribution* over possible "share" allocations.[12] For instance, we can specify a distribution over the shares of cases with ethnic homogeneity (θ^E_1), and a distribution over the shares of θ^M types, with our degrees of uncertainty captured by each distribution's variance. (More on these distributions in Chapter **??**.)

In the default setup, we assume that each θ term ($\theta^I, \theta^E, \theta^R$, etc.) is generated independently of the others.

While this is not without loss of generality, it is not as constraining as it might at first appear: any graph in which two θ terms are *not* independent can be replaced by a graph in which these two terms are themselves generated by a common, third θ term.[13] This independence feature is critical for being able to read off relations of conditional independence from a graph (see Box below). If it cannot be defended then the graph needs to be modified to communicate that θs are not independent, typically using two headed arrows. More on this in section **??**.

[11] Ordering the parent nodes alphabetically, the types we would be setting to a low probability would be $\theta^M_{0010}, \theta^M_{0110}, \theta^M_{0111}, \theta^M_{0011}$.

[12] More strictly our uncertainty is over probabilities. However it is sometimes more intuitive to describe uncertainty over shares. The distinction is not important for the applications later in which we typically assume units are independently drawn from a large population.

[13] Operationally, in the `CausalQueries` package, we can specify nodal types as having joint distributions.

Technical Note on the Markov Property

The assumptions that no node is its own descendant and that the θ terms are generated independently make the model *Markovian*, and the parents of a given node 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: nodes 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:

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

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:

$$P(v_1, v_2, \dots v_n) = \prod P(v_i|pa_i) \tag{2.2}$$

The distribution P on θ 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 nodes 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 nodes 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 nodes, 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 our setup, the θ 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 develop here. We will see how the encoding of contextual knowledge—beliefs that some kinds of conditions and causal effects 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.

2.3 Graphing models and using graphs

While we have been speaking to causal graphs throughout this chapter, we want to take some time to unpack their core features and uses. A key benefit of causal models is that they lend themselves to graphical representations; in turn, graphs constructed according to particular rules can aid causal analysis. In the next subsection we discuss a set of rules for representing a model in graphical form. The following subsection then demonstrates how access to a graph facilitates causal inference.

2.3.1 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 means 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 nodes, 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.[14] Here lies an important asymmetry: drawing an $A \rightarrow B$ 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 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 R appears somewhere in M's functional equation, as a node's functional equation must include all nodes pointing directly into it. Imagine, though, that M's causal function were specified as: $M = RE$. This function would allow for the *possibility* that R affects M, as it will whenever $E = 1$. However, it would also allow that R will have no effect, as it will when $E = 0$.

[14] 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 nodes in the model.

No excluded common causes. Any cause common to multiple nodes on a graph must itself be represented on the graph. If A and B on a graph are both affected by some third node, C, then we must represent this common cause. Thus, for instance, the graph in Figure **??** implies that I and E have no common cause. 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 node 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 nodes 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 nodes 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^J's from one another. Imagine, for instance, that the distribution of θ^D were different if $I = 0$ or $I = 0$. This would represent a classic form of confounding: the assignment of cases to values on the explanatory node would be correlated with the case's potential outcomes on D. The omission of any such pathway is precisely equivalent to expressing the belief that I is exogenous, i.e., (as if) randomly assigned.

Representing unobserved confounding. It may be however that there are common causes for nodes that we simply do not understand. We might believe that some unknown factor (partially) determines both I and D, which is the same as saying that θ^I and θ^D are not independently distributed. If we were to represent the θ terms on the graph we might then want to represent a single term (θ^I, θ^D) that points into both I and D. Usually however the θ terms are omitted from graphs and in this case we would represent the unobserved confounding by adding a dotted line, or a two headed arrow, connecting nodes whose unknown components are not independent. Figure **??** illustrates. We address this kind of unobserved confounding later in the book and show how we can seek to learn about the joint distribution of nodal types in such situations.

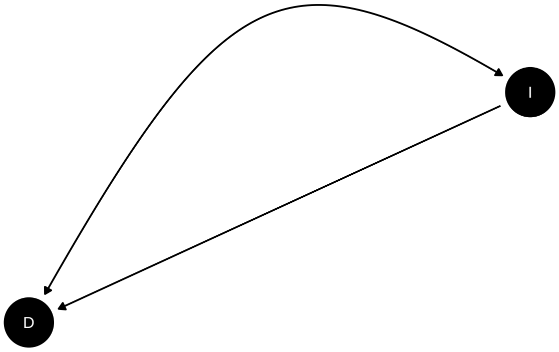


Figure 2.2: A DAG with unobserved confounding

Licence to exclude nodes. The flip side of the "no excluded common causes" 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 nodes in the model.* Indeed, θ^I in Figure **??** already implicitly captures all factors that affect I, just as θ^D captures all factors *other than* mobilization that affect democratization. We may be aware of a vast range of forces shaping 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 nodes included in the model.

We 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 nodes enter into the structural equation for every other node: 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) or interactive (the effect of each depends on the value of the other), or whether either effect is linear, concave, 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 interactive effects. As one thinks about the variety of possible causal functions, however, it quickly becomes clear that there would be no simple visual way of capturing all possible functional relations. Moreover, causal graphs do not require functional statements to perform their main analytic purpose—a purpose to which we now turn.

2.3.2 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. (For another, somewhat more extended treatment of the ideas in this section, see **?**.)

To begin thinking about conditional independence, it can be helpful to conceptualize dependencies between nodes 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 we expect I and R to be correlated. Put differently, observing the value of one of these nodes 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 we learn depends on *what we 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 and, 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 nodes, A and C, are "conditionally independent" given a set of nodes $\mathcal{B}$ 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 we learn that there have been military casualties, this information will lead us 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.

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 node, X, on a dependent node, 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

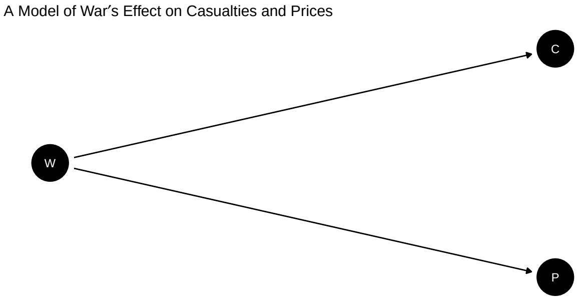


Figure 2.3: This graph represents a simple causal model in which war (W) affects both military casualties (C) and price inflation (P).

that including the covariates in the model "breaks" any biasing correlation between the value of the causal node and its unit-level effect.

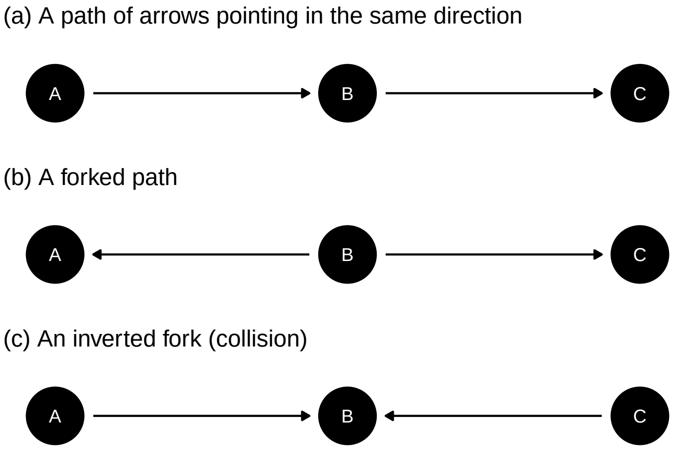
As we will explore, however, relations of conditional independence are also of more general interest in that they tell us, given a model, *when information about one feature of the world may be informative about another feature of the world, given what we already know*. By identifying the possibilities for learning, relations of conditional independence can thus guide research design. We discuss these research-design implications in Chapter **??**, but focus here on showing how relations of conditional independence operate on a DAG.

To see more systematically how a DAG can reveal conditional independencies, it is useful spell out three elemental structures according to which which information can flow across a causal graph:

(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 node along a path of arrows pointing in the same direction *blocks* flows of information across that node. 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


(a) A path of arrows pointing in the same direction

(b) A forked path

(c) An inverted fork (collision)

Figure 2.4: Three elemental relations of conditional independence.

vice-versa.

(2b) Learning the value of the node 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.[15]

(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 3 is a "collider" for A and C.[16] Although information does not flow unconditionally across colliding sequences, it does flow across them *conditional* on knowing the value of the collider node or any of its downstream consequences. In Panel 3, 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.

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

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

It is easy enough to see that, for two nodes that are correlated unconditionally, that correlation can be "broken" by controlling for a third node. In the case of collision, two nodes that are *not* correlated when taken by themselves *become* correlated when we condition on (i.e., learn the value of) a third node, 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 we already know that there was no mobilization. Now, if we observe that there *were* redistributive preferences, we 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 of any DAG. We do so by checking every path connecting two nodes of interest and ask whether, along those paths, the flow of information is open or blocked, given any other nodes whose values are already observed. Conditional independence is established when *all* paths are blocked given what we already know; otherwise, conditional independence is absent.

Following **?**, we will sometimes refer to relations of conditional independence using the concept of *d-separation.* We say that variable set $\mathcal{C}$ $d-$separates variable set $\mathcal{A}$ from variable set $\mathcal{B}$ if $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{C}$. We say that $\mathcal{A}$ and $\mathcal{B}$ are $d-$connected given $\mathcal{C}$ if $\mathcal{A}$ and $\mathcal{B}$ are *not* conditionally independent given $\mathcal{C}$.

2.3.3 Simplifying models

It is very easy to write down a model that is too complex to use effectively. In such cases we often seek simpler models that are consistent with models we have in mind but contain fewer nodes or more limited variation. In general this is possible but caution has to be taken to ensure that simplified models are indeed consistent with the original model.

Fortunately the mapping between graphs and relations of conditional independence give guidance for determining when and how it is possible to simplify models. We focus discussion on simplifications that involve node elimination and conditioning on nodes.

2.3.3.1 Eliminating nodes

If we want to eliminate a node the key rule is that the new model (and 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 if we had a model in which $A \rightarrow M \leftarrow B$ and $M \rightarrow Y$, if we were to eliminate M, M's parents (A and B) would need to adopt M's child, Y.

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.[17]

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 re-

[17] 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.

placed 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.)

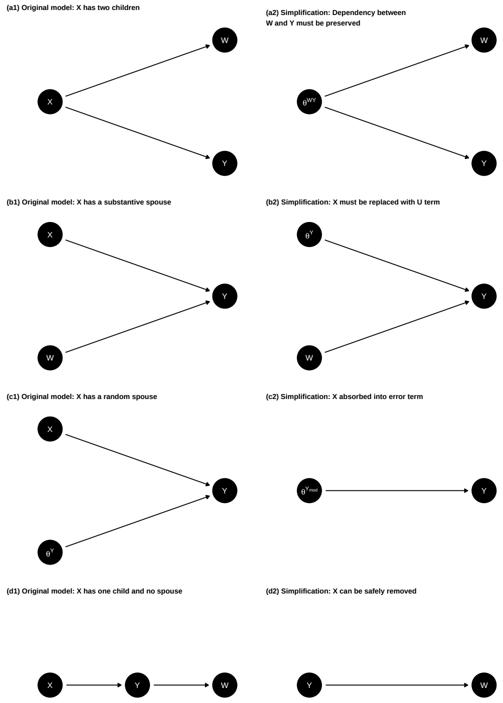

(a1) Original model: X has two children
X
W
Y
(a2) Simplification: Dependency between W and Y must be preserved
θ^WY
W
Y
(b1) Original model: X has a substantive spouse
X
Y
W
(b2) Simplification: X must be replaced with U term
θ^Y
Y
W
(c1) Original model: X has a random spouse
X
Y
θ^Y
(c2) Simplification: X absorbed into error term
θ^Y_mod
Y
(d1) Original model: X has one child and no spouse
X
Y
W
(d2) Simplification: X can be safely removed
Y
W

Figure 2.5: Basic principles for eliminating exogenous nodes.

- *Multiple children.* In (a1), we start with a model in which X has two children, thus generating a dependency between W and Y. If we eliminate X, we must preserve this dependency. We can do so, as pictured in (a2), by replacing X with a θ term that points into both W and Y. By convention, we could, alternatively, convey the same information with a dashed, undirected line between W and Y. 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 spec-

ified, 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 θ term, θ^Y, to capture the variation in Y that is not accounted for by W.

- *θ-term spouse.* In (c1), X has a spouse that is *not* substantively specified, U^Y. Eliminating X requires, again, capturing the variance that it generates as a random input. As we already have a θ term pointing only into Y, we can substitute in θ^Y_{mod}, which represents both U^Y and the variance generated by X.[18]
- *One child, no spouse.* In (d1), X has only one child and no spouse. 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 claims from simplified models with greater probabilism from theories with greater determinism.

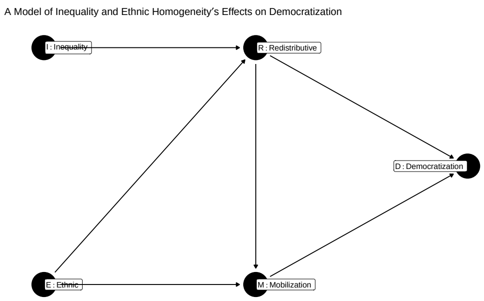


Figure 2.6: A model from which multiple simpler models can be derived.

[18] This aggregation cannot occur if θ^Y 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 original graph.

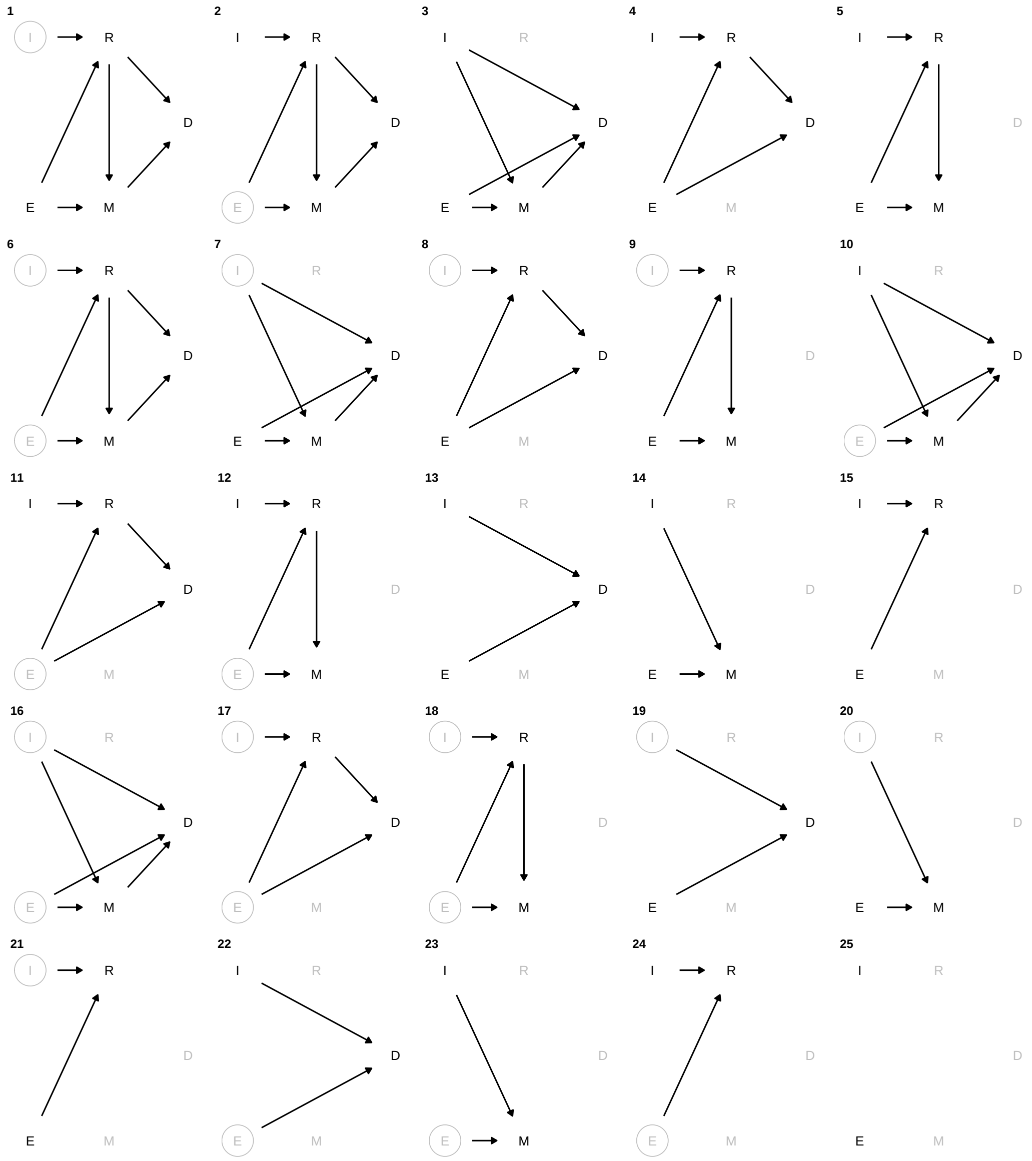


Figure 2.7: Simplifications of the model of Figure reffig:lowercomplexdem. Nodes that are eliminated are marked in grey; circles denote exogenous nodes that are replaced in subgraphs by unidentified variables. (A circled node pointing into two other nodes could equivalently be indicated as an undirected edge connecting the two.)

We can apply these principles to a model of any complexity. We illustrate a wider range of simplifications by starting with Figure **??**. 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 original model. In each subgraph,

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

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

2.3.3.2 Conditioning on nodes

Another way to simpify a model is to condition on the value of a node. When we condition on a node, we are restricting the 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 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 model $X \rightarrow M \rightarrow Y$ and we condition on the mediator, M, we sever the link between Y and 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 θ^Y.

2.3.4 Retaining probabilistic relations

We have highlighted the graphical implications of node elimination or node conditioning but importantly the distribution over θ also needs to be preserved faithfully in a move to a simpler model.

In sum, we can work with models that are simpler than our causal beliefs: we may believe a 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.

2.4 Conclusion

In this chapter, we have shown how we can inscribe causal beliefs, rooted in the potential outcomes framework, into a causal model. In doing so, we have now set out the foundations of the book's analytic framework. Causal models are both the starting point for analysis in this framework and the object about which we seek to learn. Before moving on to build on this foundation, we aim in the next chapter to offer further guidance by example on the construction of causal models, by illustrating how a set of substantive social scientific arguments from the literature can be represented in causal model form.

2.5 Chapter Appendix

Steps for constructing causal models

1. Identify a set of variables in a domain of interest. These become the nodes of the model.

- Specify the range of each node: is it continuous or discrete?
- Each node should have an associated θ term representing unspecified other influences (not necessarily graphed)

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

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

3. Write down one causal function for each endogenous node

- Each node's function must include all nodes directly pointing into it on the graph as well as the θ terms
- Functions may express arbitrary amounts of uncertainty about causal relations

4. State probabilistic beliefs about the distributions of the θs.

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

2.5.2 Model construction in code

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

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

# 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.
# You might plot it like this:
hj_ggdag(model=model)

# 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.

2.5.3 Rules for moving between levels

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 θ term. If the node has more than one child, it must be replaced with a θ 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 θ term. Where the spouse is (already) a θ term with no other children, θ 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.

2.5.4 Reading conditional independence from a graph

We illustrate how to identify the relations of conditional independence between A and D in Figure **??**.

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 node or its consequences is not known. Since no information

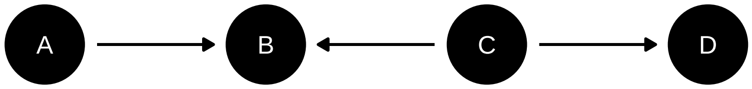


Figure 2.8: An exercise: A and D are conditionally independent, given which other node(s)?

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

Illustrating Causal Models

We use three arguments from published political science research to illustrate how to represent theoretical ideas as structural causal models.

In this short chapter, we provide more of a sense of how we can encode prior knowledge in a causal model by asking how we might construct models in light of extant scholarly works. We undertake this exercise by drawing on three well-known publications in comparative politics and international relations: Paul 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 the causal knowledge that we might plausibly think we take away from the work in question in the form of a causal model.

Readers might represent these knowledge bases differently; our aim here is only to illustrate how causal models are constructed, rather than to defend a particular representation (much less the works in question) as accurate.

For each exercise below, we focus on a specific argument in the literature in order to fix in place a relatively clear set of background causal beliefs and simplify the exposition. We emphasize, however, that *in general* a causal model should be thought of as a representation of our state of knowledge or beliefs about causal relations within a domain, rather than as a representation of a specific argument. Suppose, for instance, that we are interested in testing a specific argument in which X affects Y through the mediator M. In constructing a causal model to guide our empirical analysis, we cannot simply draw that argument in DAG form ($X \to M \to Y$) and leave it at that. In line with the principles relating to conditional independence outlined in Chapter **??**, we must consult our beliefs about this causal domain in a broader sense. For instance, given what we know about the domain from prior observations or studies, is it plausible that X could affect Y through a pathway that does not go through M? If we believe it is possible, then we must also draw a direct $X \to Y$ arrow, or our causal model will steer us wrong — even if our primary aim is to examine the pathway through M. Otherwise, our DAG will contain a relation of conditional independence (X being conditionally independent of Y given M) that we do not believe holds. Thus, while we draw on specific works in the illustrations in this chapter, we urge readers to remember that in practice one would want to characterize a broader prior knowledge base in relation to a causal domain in generating a causal model.

We aim to illuminate a number of features of causal models and their construction with these exercises. The examples that we work through variously illustrate how graphs capture beliefs about relations of conditional independence; the potential causal complexity embedded in the causal structures implied by common social-scientific arguments; and the elements of a causal model that cannot be read from a graph. For each work, we discuss both a parametric rendering of the causal functions and a non-parametric formulation built on nodal types.

3.1 Welfare state reform

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 for instance by 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 the responsibility for them difficult to trace. These blame-avoidance opportunities are themselves conditioned by the particular social-program structures that governments inherit.

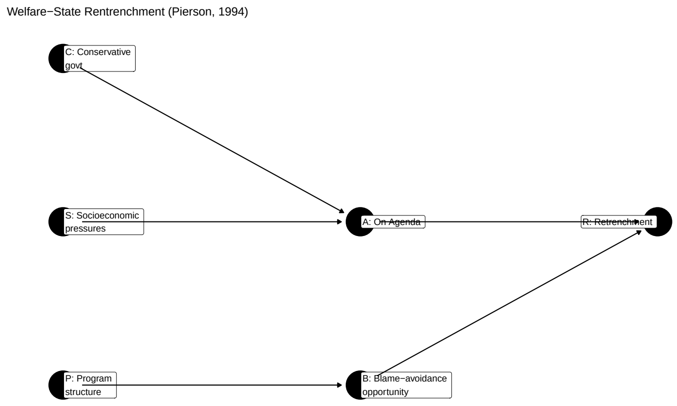

Welfare−State Rentrenchment (Pierson, 1994)
C: Conservative govt
S: Socioeconomic pressures
A: On Agenda
R: Retrenchment
P: Program structure
B: Blame−avoidance opportunity

Figure 3.1: A graphical representation of Pierson (1994).

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 makes it onto the agenda (A) and on whether blame-avoidance strategies are available to governments (B). 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. To avoid cluttering the graph, we do not represent the θ terms, but it is implied that every node on this graph has a θ node pointing into it.

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 implies that C, S, and P—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 make retrenchment politically perilous. Given the absence of bidirectional arrows indicating confounding, the graph similarly implies that the nodal types for all nodes are independent of one another.

Further, this graph represents the belief that 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 in other ways, such as by conditioning the fiscal costliness of the welfare state and thus helping to determine whether reform makes it onto the agenda. If the current state of knowledge suggested that program structures could affect retrenchment via a pathway other than blame-avoidance opportunities, then we would indeed want to include a direct $P \rightarrow A$ arrow.

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 arise, 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 T, 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.[1]

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.

[1] 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.

Let us assume that all variables (including the implied θ terms) are binary, with each condition either absent or present.

One option would be to take a parameteric approach and imagine specific functions connecting parents to children, with θ terms representing exogenous "noise." For instance, we can capture quite a lot of Pierson's theoretical logic with the following quite simple functional equations:

- $A = CS\theta^C$, capturing the idea that retrenchment makes it on the agenda only if conservatives are in power *and* socioeconomic pressures are high.
- $B = P\theta^P$, implying that blame-avoidance opportunities arise only when program structures take a particular form.
- $R = AB\theta^R$, implying that retrenchment will occur only if it is on the agenda and blame-avoidance opportunities are present.

In each equation, the θ term allows for exogenous forces that might block the outcome from occurring. In the last functional equation, for instance, retrenchment will only occur if retrenchment is on the agenda and blame-avoidance opportunities are present, but even if both are present, the effect on retrenchment also hinges on the value of θ^R. 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. We can think of θ^R as 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.

Alternatively, we could take a non-parameteric approach, as we generally do in the remainder of this book. In a non-parametric setup, each node's θ term captures that node's nodal type. Each value of a θ term's range represents a possible way in which the node might respond to its parents. We would define θ^A as taking on one of 16 values (16 types, given 2 parent nodes); θ^B as taking on on one of four values; and θ^R as taking on one of 16 values; with θ^C and θ^S each taking on one of two values.

Then choices on the probability distributions fully reflect the ways these variables relate to each other. Note that the parametric argument given above can be thought of as a special case of the non-parametric representation with all probability mass placed on a small set of possible nodal types. Thus the central thrust of Pierson's argument could then be represented in nodal-type form as:

- $\theta^A = \theta^A_{0001}$
- $\theta^B = \theta^B_{01}$
- $\theta^R = \theta^R_{0001}$.

In practice, however we would allow for a richer probability *distribution* over each θ, representing beliefs over the assignment process or causal relations operating at each node. Beliefs about the distribution of exogenous conditions would be captured in distributions over the values of θ^C, θ^S, and θ^P. How we handle distributions over θ^A, θ^B, and θ^R depends on the degree of confidence that we want to express in Pierson's argument. To represent the belief that Pierson's argument is correct with certainty and operates in uniform, deterministic fashion across units, we would simply have degenerate distributions for θ^A, θ^B, and θ^R, with a probability of 1.0 placed on the respective nodal types shown above. To capture uncertainty about the functional relations on any graph or if we believe that there is some heterogeneity of effects across units we would disperse probability density across types for each θ. For instance, for θ^R we might want to put some weight on θ^R_{0011} (blame-avoidance opportunities alone are enough to generate retrenchment), θ^R_{0101} (conservative leaders alone are enough), θ^R_{0111} (either is enough), and θ^R_{0000} (retrenchment will not happen even when both conditions are present), while perhaps putting greatest weight on θ^R_{0001}.[2]

3.2 Military Interventions

? 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

[2] In notation that we use later, these beliefs would be represented with a λ^R vector.

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 leaders make about intervention strategies—in particular, about whether to try to transform the internal institutions of a target state—via two mechanisms. First, an internal focus (as opposed to an external focus) affects leaders' perceptions of the likely security gains from a transformative intervention strategy. Second, internal vs. external focus affects the kinds of strategic capabilities in which leaders invest over time (do they invest in the kinds of capabilities suited to internal transformation?); and those investments in turn affect 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.

At the same time, leaders can only choose a transformative strategy if they decide to intervene at all. The decision about whether to intervene depends, in turn, on at least two kinds of considerations. The first is about fit: a leader is more likely to intervene against a target when the nature of the dispute makes the leader's preferred strategy appear feasible in a given situation. Second, Saunders allows that forces outside the logic of her main argument might also affect the likelihood of intervention: in particular, leaders may be pushed to intervene by international or domestic audiences.

Figure **??** depicts the causal dependencies in Saunders' argument in DAG form (again, with all θ terms implied). Working from left to right, we see that whether or not leaders are "internally focused" (F) affects the expected net relative benefits of transformation (B), both via a direct pathway and via an indirect pathway running through investments in transformative capacities (T). Characteristics of a given dispute or target state (D) likewise influence the benefits of transformation (B). The decision about whether to intervene (I) is then a function of three factors: internal focus (F), the expected relative net benefits of transformation (B), and audience pressures (A). Finally, the choice of whether to pursue a transformative strategy (S) is a function of whether or not intervention occurs at all (I), and of cost-benefit comparisons between the two strategies (B).

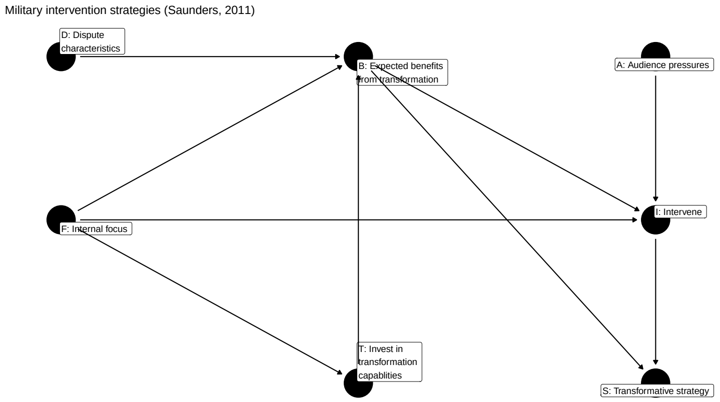

Military intervention strategies (Saunders, 2011)
D: Dispute characteristics
B: Expected benefits from transformation
A: Audience pressures
F: Internal focus
I: Intervene
T: Invest in transformation capablities
S: Transformative strategy

Figure 3.2: A graphical representation of Saunders' (2011) argument.

This DAG illustrates how readily causal graphs can depict the multiple pathways through which a given variable might affect another variable, as with the multiple pathways linking F 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 F to the outcomes of interest, I and S. What is a bit easier to miss without formalization is that F 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 D and A suggests that features of the dispute that affect feasibility have no effect on audience pressures. If we instead believed there could be other connections—for instance, that audiences take feasibility into account in demanding intervention—then we would want to include a $D \to A$ 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, $F = 1$ representing internally focused causal beliefs, $T = 1$ representing investments in transformative capabilities, $B = 1$ representing expectations that transformation will be more net beneficial than non-transformation, $D = 1$ meaning that a dispute 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:

- $B = FTD$, 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 dispute has characteristics that make transformation feasible.
- $I = (1-|B-F|)+(1-(1-|B-F|))A$, 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 $B - F = 0$) or, when this alignment is absent (i.e., such that $|B - F| = 0$), where audiences pressure a leader to intervene.

As illustrated in the Pierson example, in a non-parametric framework, each parametric functional equation represents one nodal type for the relevant θ. For instance, though we spare the reader the complexities of the corresponding subscript notation, there is a single value of θ^B under which the conditions $F = 1, P = 1$, and $T = 1$ generate $B = 1$, and we get $B = 0$ otherwise. Likewise, there exists a single value of θ^I under which $B = 1, F = 1$ and $B = 0, F = 0$ produce $I = 1$, for either value of A; and A has a positive effect on I whenever $B \neq F$. To work with this model, we would specify a probability distribution over all possible nodal types for each node on the graph.

This example also nicely illustrates how much potential causal complexity a moderately intricate argument and causal graph implies. The number of *possible* nodal types at each node depends on how many parents that node has. Looking at the endogenous nodes here, we have one node with one parent (T), implying 4 nodal types; one node with two parents (S), implying

16 nodal types; and two nodes with 3 parents (B and I), implying 256 nodal types each. If we now conceptualize the set of possible "causal types" as containing all distinct combinations of nodal types—all ways in which a case might behave across all of its nodes (see Chapter **??**)—then this graph implies about 4 million different ways in which the values of exogenous nodes (D, F, and A) might jointly produce patterns of outcomes. Saunders' argument effectively represents one of these 4 million possible sets of relations.

The framework that we outline in this book allows for updating on arguments like Saunders': we can ask how likely the specific causal type implied by this argument is relative to other causal types. Yet, as we will see, the approach lends itself to a much broader view of causal inquiry. In general, we will use data to update beliefs over *all* causal types allowed for in a model, and then use these updated beliefs to answer any number of causal questions about relations in the model. For instance, we can use the same data and updated model to ask about the average effect of internal focus on intervention; the relative importance in this effect of the expected-benefits pathway over the direct pathway; about individual steps in the causal chain, such as the effect of expected benefits on choice of strategy; and so on.

3.3 Development and Democratization

? 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, in their description—and contrary to **?** —democratization is "exogenous": it is not determined by other variables in the model. 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}, not pictured).

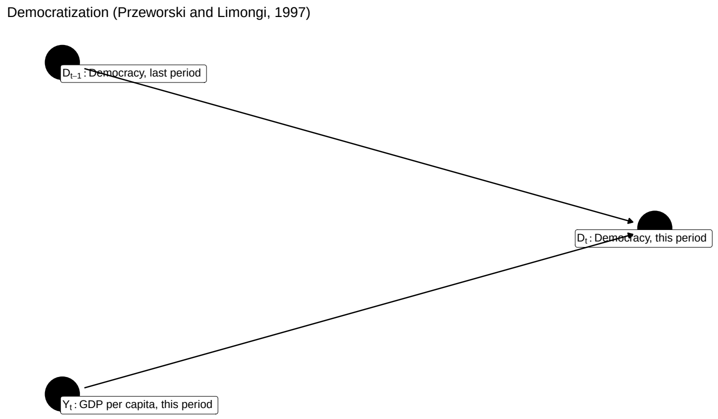

Democratization (Przeworski and Limongi, 1997)
D_{t-1} : Democracy, last period
D_t : Democracy, this period
Y_t : GDP per capita, this period

Figure 3.3: 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.

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 absence 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.

Inspection of this figure highlights, we think, a curious feature of this argument. The key claim—that the switch *to* democracy does not depend on income—is not readable from the graph. The reason is simply that, given this argument, being a non democracy in one period given you were a democracy in a previous period does depend on income. Your state in the second period does depend on income. Specifically there is a *counterfactual* dependence—income causes a state to be democratic even if it does not cause it to *transition to* democracy. The effect of income may well be asymmetric depending on whether you start as a democracy or start as an autocracy but this asymmetry has to be be captured by the specification of the functional relations; it is not captured by the graph.

For a parametric representation of this asymmetric relationship we can specify a function in which *GDP* can reduce the likelihood of a transition *away* from democracy but does affect the probability of a transition *to* democracy, which should be exogenously determined. One possible translation of the argument into functional terms is:

$$D_t = \mathbb{1}(\theta^{D_t} + (1 - D_{t-1})p + D_{t-1}Y_{t-1}q) > 0)$$

where

- D_t and D_{t-1} are binary, representing current and last-period democracy, respectively
- p is a parameter representing the probability that an autocracy democratizes
- q is a parameter representing the probability that a democracy with a GDP of 1 remains democratic
- Y_{t-1} represents national per capita GDP, scaled to 0-1.
- θ^{D_t} represents a random, additional input into democracy
- 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 inequality 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 q times Y_{t-1}: thus, remaining democratic is more likely as national income rises. The inequality is then evaluated by asking whether the expression on the left passes a threshold. 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. The mean and variance of θ^{D_t} capture the overall likelihood of being a democracy as well as the importance of unspecified factors.[3] In a model like this it would be natural to seek to estimate parameters p and q as well as trying to understand the distribution of θ^{D_t}.

We can also represent the asymmetry in the binary set up with causal types that we developed in the last chapter.

Type $\theta^{D_t}_{0001}$ is a type for which the regime type *will stay as they are* if they are wealthy, but will become authoritarian if they are not wealthy. To be clear, wealth still affects whether or not a state is a democracy, rather than an autocracy, in this period–counterfactually—but wealth does not make a non democracy become a democracy. In other words it causes a case to *be* a democracy but not to *become* a democracy. This type can be distinguished from a $\theta^{D_t}_{0011}$ type in which a non democracy becomes a democracy when income is high.

Although we do not engage with dynamic models in this book, it is instructive to think through the implications of a distribution of causal types for a dynamic process. Say we were to imagine that income were constant but that in each period one half of units were of type $\theta^{D_t}_{0001}$, and one half of type $\theta^{D_t}_{1111}$ (with types drawn afresh each period). Say that in an initial period, half the units were democracies and half had high income and there was no relation between these two features. Then in the next period we would have that half of cases would be democracies (regardless of income), half of which would be surviving democracies and half new democracies; of the other half, one quarter would be surviving democracies (surviving *because of* their income), and the other three quarters would be autocracies, one third of which would be "backsliders" because of their poverty. Similar transitions occur in future periods until eventually the wealthy states are all stable democracies and the poorer states transition between democracy and autocracy back and forth randomly each period.

In this approach there are no parameters p or q to be estimated. Rather the

[3] 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 y into a change in the probability of reversion or allowing for non-linearities, with unknown parameters, in this effect.

focus is entirely on the distribution of nodal types.

Chapter 4

Causal Queries

We describe major families of causal queries and illustrate how these can all be described as queries about the values of nodes in a causal model.

Although scholars share a broad common interest in causality there is tremendous heterogeneity in the kinds of causal questions that scholars ask. Consider the relationship between inequality and democratization. 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 inequality would have an effect *here*. That is a question about causal effects at the case level. Alternatively we might wonder whether the level of democracy in Mongolia in 1995 is *due* to the level of inequality in that case—this is quite a distinct question (in the same way that establishing that poison would make you sick does not imply that you are sick because of poison). Or 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; or we can ask how commonly inequality affects democratization through mobilization across a broad set of cases. Pushing further we might ask a counterfactual question of the form: would inequality have produced democratization had mobilization been prevented from occurring.

Distinct 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. Fortunately each of these questions can be readily captured as specific queries asked of (and answerable from) a causal model. As described by **?**, the goal is to deploy an "*algorithm* that receives a model *M* as an input and delivers the desired quantity *Q(M)* as the output." More specifically, we demonstrate how, given the structure we described in Chapter **??**, 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 in no way exhaustive of the causal questions that can be captured in causal graphs, but they are among the more common social scientific investigations.

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 intervene to change X's value in the case, (how) would Y's value 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 setup with binary variables 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 nodal type characterizing Y's response to X and, if X and Y are binary, it can take on one of four values: θ^Y_{10}, θ^Y_{01}, θ^Y_{00}, and θ^Y_{11} (which map onto our a, b, c and d types, respectively). 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 $\theta^Y = \theta^Y_{01}$), and 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 the nodal type* θ^Y.

Two natural variants of this question are, "What is the expected effect of X on Y?" and "What is the probability that X matters for Y? Answering the question requires estimating the probability that X has a positive effect minus the probability that it has a negative effect: $\Pr(\theta^Y = \theta^Y_{01}) - \Pr(\theta^Y = \theta^Y_{10})$. Answering the second involves assessing $\Pr(\theta^Y = \theta^Y_{01} \text{ OR } \theta^Y = \theta^Y_{10})$.

We can conceptualize this same question even if the model involves more complex relations between X and Y. The question itself does not depend on the model having a particular form. For instance, consider a mediation model of the form $X \rightarrow M \rightarrow Y$. In this model, a positive effect of X on Y

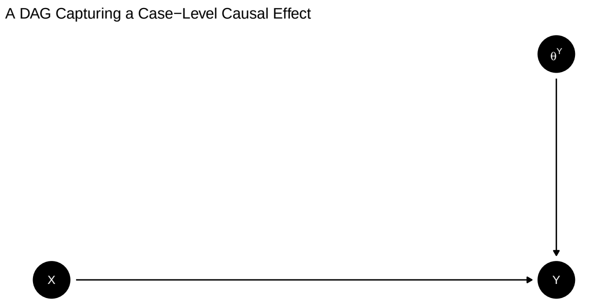


Figure 4.1: This DAG is a graphical representation of the simple causal setup in which the effect of X on Y in a given case depends on the case's nodal type, represented by θ^Y. With a single binary causal variable of interest, we let θ^Y take on values θ^Y_{ij}, with i representing the value Y takes 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}.

can emerge either from a chain of positive effects of X on M and of M on Y or from a chain of negative effects, while a negative effect of X on Y can emerge from a chain of opposite-signed effects. Thus, answering the question means estimating:

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

Answering the question now requires guesses about nodal types for M and for Y, not just nodal types for Y. Thus the question remains the same, and answerable, under different models, but the answer to the question might involve summaries of the values of different nodes.

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 asks: "*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? Or: 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 the θ nodes on endogenous variables, 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_1 and X_2. 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 all possible combinations of X_1 and X_2, including interactions between them.

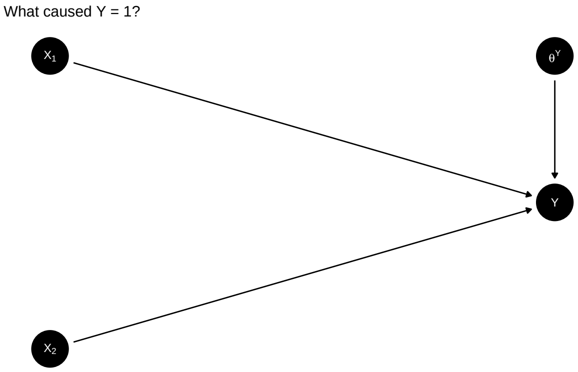


Figure 4.2: This DAG is a graphical representation of the simple causal setup in which Y depends on two variables $X1$ and $X2$. How Y responds to X1 and X2 depnds on θ^Y, the DAG itself does not provide information on whether or how X1 and X2 interact with each other.

We examined the set of nodal 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 X_1 and X_2 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 nodal-type "space" is to note that X_1 can have any of our four causal effects (the four binary types) on Y when $X_2 = 0$; and X_1 can have any of four causal effects when $X_2 = 1$.[1] This yields 16 possible response patterns to combinations of X_1 and X_2 values.

A query about causal attribution—whether $X_1 = 1$ caused $Y = 1$—for the model in Figure **??**, needs to be defined in terms of both X_2 and θ^Y. Parallel to our Taiwan example, suppose that we have a case in which $Y = 1$ and in which X_1 was also 1, and we want to know whether X_1 caused Y to take on the value it did. Answering this question requires knowing whether the case's type is such that X_1 would have had a positive causal effect on Y, *given the value of* X_2 (which we can think of as part of the context). Thus, given that we start with knowledge of X_1's and Y's values, our query about causal attribution amounts to a query about two nodal types: (a) θ^{X_2} (which gives X_2's value) and (b) θ^Y, specifically whether its value is such that X_1 has a positive causal effect given X_2's value.

Suppose, for instance, that we were to observe $X_2 = 1$. We then need to ask whether the nodal type, θ^Y, is such that X_1 has a positive effect when $X_2 = 1$. Consider θ^Y_{0111} (type 8 in Table **??**).[2] This is a nodal type in which X_1 has a positive effect when $X_2 = 0$ but no effect when $X_2 = 1$. Put differently, $X_2 = 1$ is a sufficient condition for $Y = 1$, meaning that X_1 makes no difference to the outcome when $X_2 = 1$.

In all we have four qualifying Y-types: θ^Y_{0001}, θ^Y_{1001}, θ^Y_{0101}, θ^Y_{1101}. In other words, we can attribute a $Y = 1$ outcome to $X_1 = 1$ when $X_2 = 1$ and Y's nodal type is one of these four.

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 the nodal type(s) governing effects is such that changing the causal factor of interest would have changed the outcome.

[1] This is precisely equivalent to noting that X_2's effect on Y can be of any of the four types when $X_1 = 0$ and of any of the four types when $X_1 = 1$.

[2] A reminder that, with two-parent nodes, the nodal-type subscript ordering is $Y|(X_1 = 0, X_2 = 0); Y|(X_1 = 1, X_2 = 0); Y|(X_1 = 0, X_2 = 1); Y|(X_1 = 1, X_2 = 1)$.

4.3 Actual causes

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 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* can be useful in putting a finer point on this kind of causal question.

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 excellent 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 (say) the economic crisis occurring first and toppling the government before the war could do so. In this situation, neither economic crisis nor war in fact made a difference to the outcome: take away either one and the outcome remains the same.

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. We can formalize this intuition by defining Sally's throw as the *actual cause* of the outcome. 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).

Let us now qpply these 3 conditions to the Sally and Billy example. 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 — with the important caveat that, 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: (1) Sally and Billy throw, (2) Sally's or Billy's rock hits the bottle, and (3) the bottle breaks. In actuality, Billy's stone did not hit the bottle, so we are allowed to condition on that fact in determining whether Sally's throw was a counterfactual cause. Conditioning on 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 in that sense.[3]

As with other causal queries, the question "Was $X = x$ the actual cause of $Y = y$?" can be redefined as a question about which combinations of nodal types 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 nodal type. We capture the possible "preemption" effect with the arrow pointing from H^S to H^B, allowing whether Sally's rock hits to affect whether Billy's rock hits.[4]

For Sally's throw to be an actual cause of the bottle's cracking, we need first to establish that Sally threw ($\theta^S = \theta^S_1$) and that the bottle cracked ($\theta^C = \theta^C_1$) (Condition 1). Condition 3 is automatically satisfied in that $\theta^S = \theta^S_1$ has no strict subsets. Turning now to Condition 2, we need Sally's throw to be a counterfactual cause of the bottle cracking if we condition on the value of some set of nodes remaining fixed at the values they in fact took on. As discussed above, we know that we can meet this criterion if we condition on

[3] Perhaps more surprising, it is possible that the expected causal effect is negative but that X is an actual cause in expectation. For instance, suppose that 10% of the time Sally's shot intercepts 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 **?**.

[4] We do not need an arrow in the other direction because Sally throws first.

Billy's throw not hitting. To make this work, we need to ensure, first, that Sally's throw hits if and only if she throws: so $\theta^{H^S} = \theta^{H^S}_{01}$. Next, we need to ensure that Billy's throw does not hit whenever Sally's does: this corresponds to any of the four nodal types for H^B that take the form $\theta^{H^B}_{xx00}$, meaning that $H^B = 0$ whenever $H^S = 1$. Note that the effect of Billy throwing on Billy hitting when Sally has *not* thrown—the first two terms in the nodal-type's subscript—does not matter since we have already selected a value for θ^S such that Sally does indeed throw.

Finally, we need θ^C to take on a value such that H^S has a positive effect on C when $H^B = 0$ (Billy doesn't hit) since this is the actual circumstance on which we will be conditioning. This is satisfied by any of the four nodal types of the form θ^C_{0x1x}. 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^C_{0011}. Here, Sally's throw is a counterfactual cause of the bottle's cracking. And, as we have said, all counterfactual causes are actual causes. They are, simply, counterfactual causes when we hold *nothing* fixed ($\mathcal{W}$ in Condition 2 is just the empty set).

Notably, we do not need to specify the nodal type for B: given the other nodal types identified, Sally's throw will be the actual cause regardless of whether or not Billy throws. If Billy does not throw, then Sally's throw is a simple counterfactual cause (given the other nodal types).

The larger point is that actual cause queries can, like all other causal queries, be defined as questions about the values of nodes in a causal model. When we pose the query, was Sally's throw an actual cause of the bottle cracking, we are in effect asking whether the case's combination of nodal types (or its causal type) matches $\theta^S_1, \theta^B_x, \theta^{H^B}_{xx00}, \theta^{H^S}_{01}, \theta^C_{0x1x}$.

Likewise, if want to ask *how often* Sally's throw is an actual cause, in a population of throwing rounds, we can address this query as a question about the joint *distribution* of nodal types. We are then asking how common the qualifying combinations of nodal types are in the population given the distribution of types at each node.

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

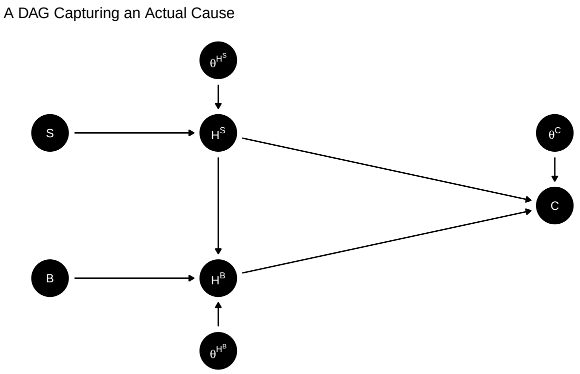


Figure 4.3: This DAG is a graphical representation of the simple causal setup in which the effect of X on Y in a given case depends on the case's nodal type for Y, represented by θ^Y.

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 distinctively contributed to the outcome, even if it was not a counterfactual cause.

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." We can say that 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 them, 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.

How notable a counterfactual cause is can be expressed as a claim about the distribution of a set of nodal types. For instance, if we observe $R^j = 1$ for state j (it voted Republican), then the notability of this vote directly increases in our belief about the probability that $\theta^{R^j} = \theta_0^{R^j}$: the probability that the state's vote could have gone the other way.

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 nodal types. When we seek to estimate an average causal effect, we seek information about the *shares* of these nodal types in the population.

More formally and adapted from **?**, we can use λ^Y_{ij} to refer to the *share* of cases in a population that has nodal type θ^Y_{ij}. Thus, given our four nodal types in a two-variable binary setup, λ^Y_{10} is the proportion of cases in the population with negative effects; λ_{01} is the proportion of cases with positive effects; and so on. One nice feature of this setup, with both X and Y as binary, is that the average causal effect can be simply calculated as the share of positive-effect cases less the share of negative-effect cases: $\lambda^Y_{01} - \lambda^Y_{10}$.

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 nodal 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 nodal 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 $\lambda^Y_{10}, \lambda^Y_{01}, \lambda^Y_{00}, \lambda^Y_{11}$) and the random process drawing a θ^Y value from that distribution as U^θ.

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^Y_{10} and λ^Y_{01}? 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?[5]

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 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. Here we have nodal types defining M's response to X (θ^M) and defining Y's response to both X (directly) and M (θ^Y).

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

[5] 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.

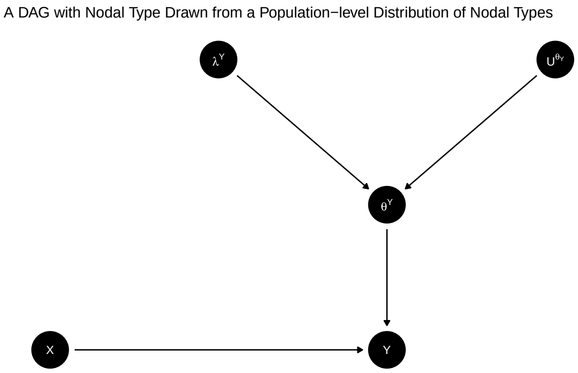

A DAG with Nodal Type Drawn from a Population−level Distribution of Nodal Types
λY
Uθy
θY
X
Y

Figure 4.4: This DAG is a graphical representation of a causal setup in which cases are drawn from a population composed of different nodal 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 nodal 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 nodal type, θ^Y, is thus a joint function of λ^Y and U^{θ^Y}.

effect ran *through* M. We will refer to this as a query about a causal path. Importantly, a causal path query is not satisfied simply by asking 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 mediated (via M) 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—that is, *will not operate without*—the effect of X on M.[6] 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.

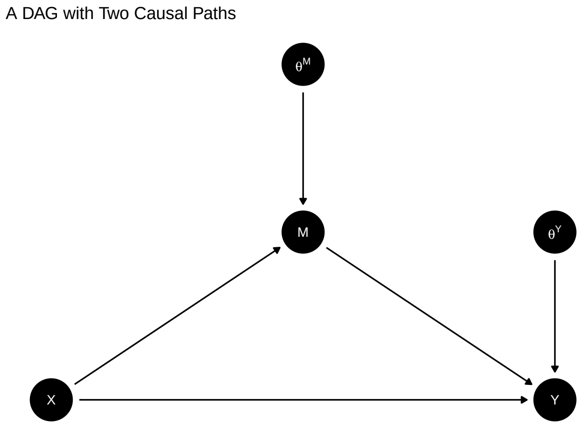


Figure 4.5: Here X has effects on Y both indirectly through M and directly.

As we can show, we can define this causal-path query as a question about specific nodes on a causal graph. In particular, a causal path can be defined

[6] 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, **?**).

in terms of the values of θ nodes: specifically, in the present example, in terms of θ^M and θ^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 θ^M. We know that θ^M can take on four possible values corresponding to the four possible responses to X: $\theta^M_{10}, \theta^M_{01}, \theta^M_{00}, \theta^M_{11}$. 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 defines Y's response to different combinations of two other variables—here, X and M—since *both* of these variables point directly into Y. Where X can have both a mediated effect through M and a direct effect, X and M also potentially *interact* in affecting 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. 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 nodal types that capture two kinds of counterfactual causal relations operating on nodes. First, X must have a positive effect on Y when M changes as it does as a result of 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 a situation in which X changes from 0 to 1,[7] but M does *not* change to the

[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

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 only if Y would *not* change to 1 in this situation that we have identified a nodal type for which the M-mediated path matters. It is this thought experiment that isolates the causal significance of the path that runs through M.

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

1. Is $X = 1$ a counterfactual cause of $Y = 1$, given X's positive effect on M? 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, 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 $X = 1$ and $M = 0$, and we ask: does $Y = 0$? Of the four remaining types, only 2 and 6 pass this test.

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 $X = 1$.

Under these and only these two values of θ^Y—θ^Y_{0001} and θ^Y_{0101}—we will see a positive effect of X on Y for which the M-mediated path is causally necessary, given a positve effect of X on M. These two θ^Y values are also different from one another in an interesting way. For type θ^Y_{0101}, 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^Y_{0101}, then, anything else that similarly affects M would generate the same effect on Y as X does. In type θ^Y_{0001}, 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. And here X itself matters in the counterfactual sense; for a case of type θ^Y_{0001}, 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_{0010} and θ_{1010} 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 ($\theta^Y = \theta_{0001}$ or θ_{0101})
- $\theta^M = \theta^M_{10}$ and ($\theta^Y = \theta_{0010}$ or θ_{1010})

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.

4.6 General procedure

We have been able to associate a collection of causal types to each of the causal queries we have described in this chapter. But we have not described a general method for doing so. We do that now.[9]

The algorithm calculates the full set of outcomes on all nodes, given each possible causal type and a collection of controlled conditions ("`do` operations"). Then each causal type is marked as satisfying the query or not. This in turn then tells us the *set* of types that satisfy a query. Quantitative queries, such as the probability of a query being satisfied, or the average treatment effect, can then be calculated by taking the measure of the set of causal types that satisfies the query.

First some notation.

Let n denote the number of nodes. Label the nodes $V_1, \dots V_n$ subject to the requirement that each node's parents precede it in the ordering. Let pa_j denote the set of values of the parents of node j and let $V_j(pa_j, \theta_t)$ denote the value of node j given the values of its parents and the causal type θ_t.

The primitives of a query are questions about the values of outcomes, V, given some set of controlled operations x.

- let $x = (x_1, \dots x_n)$ denote a set of `do` operations where each x_i takes on a value in $\{-1, 0, 1\}$. here -1 indicates "not controlled", 0 means set to 0 and 1 means set to 1 (this set can be expanded if V is not binary)
- let $V(x, \theta_t)$ denote the values V (the full set of nodes) takes given θ_t
- a "simple query" is a function $q(V(x, \theta_t))$ which returns TRUE if $V(x, \theta_t)$ satisfies some condition and FALSE otherwise.

Queries are summaries of simple queries. For instance, for nodes X and Y:

[9] In particular we describe the algorithm used by the `CausalQueries` package. This approach is not the efficient but it is intuitive and can be used for arbitrarily complex queries.

- Query $Q_1 : \mathbb{1}(Y(X = 1) = 1))$ asks whether $Y = 1$ when X is set to 1. This requires evaluating one simple query.
- Query $Q_2 : \mathbb{1}(Y(X = 1) = 1)\&\mathbb{1}(Y(X = 0) = 0))$ is composed of two simple queries: the first returns true if Y is 1 when X is set to 1, the second returns true if Y is 0 when X is set to 0; both conditions holding corresponds to a positive effect on a unit.
- Query $Q_3 : E((\mathbb{1}(Y(X = 1) = 1)\&(Y(X = 0) = 0)) - (\mathbb{1}(Y(X = 1) = 0)\&\mathbb{1}(Y(X = 0) = 1))$ asks for the average treatment effect, represented here using four simple queries: the expected difference between positive and negative effects. This query involves weighting by the probability of the causal types.

Then to calculate $V(x, \theta_t)$:

1. Calculate v_1, the realized value of the first node, V_1, given θ_t. This is given by $v_1 = x_1$ if $x_1 \neq -1$ and by $\theta_t^{V_1}$ otherwise.
2. For each $j \in 2...n$ calculate v_j using either $v_j = x_j$ if $x_j \neq -1$ and $V_j(pa_j, \theta_t)$ otherwise, where the values in pa_j are determined in the previous steps.

We now have the outcomes, V, for all nodes given the operations x and so can determine $q(V(x))$. From there we can calculate summaries of simple queries across causal types.

A last note on conditional queries. Say we are interested in an attribution query of the form: what is the probability that X causes Y in a case in which $X = 1$ and $Y = 1$. In this case define simple query q_1 which assesses whether X causes Y for a given θ_t and simple query q_2 which assesses whether $X = 1$ and $Y = 1$ under θ_t. We then calculate the conditional query by conditioning on the set of θs for which q_2 is true and evaluating the share of these for which q_2 is true (weighting by the probability of the causal types).

4.7 Chapter Appendix

We demonstrate how queries are calculated using the `CausalQueries` package for a chain model of the form $X \to M \to Y$. We imagine a model of this form in which we assume no negative effects of M on X or M on Y. We will also suppose that in fact $X = 1$, always. Doing this keeps the parameter space a little smaller for this demonstration but also serves to demonstrate that a causal model can make use of the counterfactual possibility that a node takes on a particular value even if it never does in fact.

We then ask two questions:

- Q1. What is the probability that X causes Y? ("POS")
- Q2. What is the probability that X causes Y in cases in which $X = 1$ and $Y = 1$? ("POC")

To answer these two queries we define simple query q_1 which assesses whether X causes Y for each θ and a second simple query q_2 which assesses whether $X = 1$ and $Y = 1$ for each θ. In this example the first simple query involves some `do` operations, the second does not.

Code to answer these two simple queries is shown below and the output is shown in Table **??** (one row for each causal type).

```
model <- make_model("X -> M -> Y") %>%
         set_restrictions("X[]==0") %>%
         set_restrictions("M[X=1] < M[X=0]") %>%
         set_restrictions("Y[M=1] < Y[M=0]")

q1 <- "Y[X = 1] > Y[X = 0]"
q2 <- "X == 1 & Y == 1"

df <- data.frame(
  a1 = CausalQueries:::map_query_to_causal_type(model, q1)$types,
  a2 = CausalQueries:::map_query_to_causal_type(model, q2)$types,
  p  = get_type_prob(model))
```

The answer to the overall queries are then (1) the expected value of (the answers to) q_1 and weights $p$and (2) the expected value of (the answers to) q_1 given q_0 and weights p. See Table **??**.

```
df %>% summarize(POS = weighted.mean(a1, p),
                 POC = weighted.mean(a1[a2], p[a2]))
```

Given the equal weighting on causal types, these answers reflect the fact that for 5 of 9 causal types we expect to see $X = 1$ and $Y = 1$ but that; the

Table 4.1: Set of causal types in the model that satisfy q1 and q2 along with the probability of the type.

	a1	a2	p
X1.M00.Y00	FALSE	FALSE	0.111
X1.M01.Y00	FALSE	FALSE	0.111
X1.M11.Y00	FALSE	FALSE	0.111
X1.M00.Y01	FALSE	FALSE	0.111
X1.M01.Y01	TRUE	TRUE	0.111
X1.M11.Y01	FALSE	TRUE	0.111
X1.M00.Y11	FALSE	TRUE	0.111
X1.M01.Y11	FALSE	TRUE	0.111
X1.M11.Y11	FALSE	TRUE	0.111

Table 4.2: Calculated answers to two queries.

POS	POC
0.111	0.2

causal effect is present for only 1 of 9 causal types and for 1 of the 5 causal types that exhibit $X = 1$ and $Y = 1$.

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 queries 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.