THE BOOK OF WHY

THE NEW SCIENCE OF CAUSE AND EFFECT

JUDEA PEARL
AND DANA MACKENZIE

BASIC BOOKS
New York

CONTENTS

To Ruth

BASIC
BOOKS

PREFACE

ALMOST two decades ago, when I wrote the preface to my book *Causality* (2000), I made a rather daring remark that friends advised me to tone down. "Causality has undergone a major transformation," I wrote, "from a concept shrouded in mystery into a mathematical object with well-defined semantics and well-founded logic. Paradoxes and controversies have been resolved, slippery concepts have been explicated, and practical problems relying on causal information that long were regarded as either metaphysical or unmanageable can now be solved using elementary mathematics. Put simply, causality has been mathematized."

Reading this passage today, I feel I was somewhat shortsighted. What I described as a "transformation" turned out to be a "revolution" that has changed the thinking in many of the sciences. Many now call it "the Causal Revolution," and the excitement that it has generated in research circles is spilling over to education and applications. I believe the time is ripe to share it with a broader audience.

This book strives to fulfill a three-pronged mission: first, to lay before you in nonmathematical language the intellectual content of the Causal Revolution and how it is affecting our lives as well as our future; second, to share with you some of the heroic journeys, both successful and failed, that scientists have embarked on when confronted by critical cause-effect questions.

Finally, returning the Causal Revolution to its womb in artificial intelligence, I aim to describe to you how robots can be constructed that learn to communicate in our mother tongue—the language of cause and effect. This new generation of robots should explain to us why things happened, why they responded the way

they did, and why nature operates one way and not another. More ambitiously, they should also teach us about ourselves: why our mind clicks the way it does and what it means to think rationally about cause and effect, credit and regret, intent and responsibility.

When I write equations, I have a very clear idea of who my readers are. Not so when I write for the general public—an entirely new adventure for me. Strange, but this new experience has been one of the most rewarding educational trips of my life. The need to shape ideas in your language, to guess your background, your questions, and your reactions, did more to sharpen my understanding of causality than all the equations I have written prior to writing this book.

For this I will forever be grateful to you. I hope you are as excited as I am to see the results.

Judea Pearl
Los Angeles, October 2017

INTRODUCTION: MIND OVER DATA

Every science that has thriven has thriven upon its own symbols.

—Augustus de Morgan (1864)

THIS book tells the story of a science that has changed the way we distinguish facts from fiction and yet has remained under the radar of the general public. The consequences of the new science are already impacting crucial facets of our lives and have the potential to affect more, from the development of new drugs to the control of economic policies, from education and robotics to gun control and global warming. Remarkably, despite the diversity and apparent incommensurability of these problem areas, the new science embraces them all under a unified framework that was practically nonexistent two decades ago.

The new science does not have a fancy name: I call it simply "causal inference," as do many of my colleagues. Nor is it particularly high-tech. The ideal technology that causal inference strives to emulate resides within our own minds. Some tens of thousands of years ago, humans began to realize that certain things cause other things and that tinkering with the former can change the latter. No other species grasps this, certainly not to the extent that we do. From this discovery came organized societies, then towns and cities, and eventually the science- and technology-based civilization we enjoy today. All because we asked a simple question: Why?

Causal inference is all about taking this question seriously. It posits that the human brain is the most advanced tool ever devised for managing causes and effects. Our brains store an incredible amount of causal knowledge which, supplemented by data, we

could harness to answer some of the most pressing questions of our time. More ambitiously, once we really understand the logic behind causal thinking, we could emulate it on modern computers and create an "artificial scientist." This smart robot would discover yet unknown phenomena, find explanations to pending scientific dilemmas, design new experiments, and continually extract more causal knowledge from the environment.

But before we can venture to speculate on such futuristic developments, it is important to understand the achievements that causal inference has tallied thus far. We will explore the way that it has transformed the thinking of scientists in almost every data-informed discipline and how it is about to change our lives.

The new science addresses seemingly straightforward questions like these:

- How effective is a given treatment in preventing a disease?
- Did the new tax law cause our sales to go up, or was it our advertising campaign?
- What is the health-care cost attributable to obesity?
- Can hiring records prove an employer is guilty of a policy of sex discrimination?
- I'm about to quit my job. Should I?

These questions have in common a concern with cause-and-effect relationships, recognizable through words such as "preventing," "cause," "attributable to," "policy," and "should I." Such words are common in everyday language, and our society constantly demands answers to such questions. Yet, until very recently, science gave us no means even to articulate, let alone answer, them.

By far the most important contribution of causal inference to mankind has been to turn this scientific neglect into a thing of the past. The new science has spawned a simple mathematical language to articulate causal relationships that we know as well as those we wish to find out about. The ability to express this information in mathematical form has unleashed a wealth of powerful and principled methods for combining our knowledge

with data and answering causal questions like the five above.

I have been lucky to be part of this scientific development for the past quarter century. I have watched its progress take shape in students' cubicles and research laboratories, and I have heard its breakthroughs resonate in somber scientific conferences, far from the limelight of public attention. Now, as we enter the era of strong artificial intelligence (AI) and many tout the endless possibilities of Big Data and deep learning, I find it timely and exciting to present to the reader some of the most adventurous paths that the new science is taking, how it impacts data science, and the many ways in which it will change our lives in the twenty-first century.

When you hear me describe these achievements as a "new science," you may be skeptical. You may even ask, Why wasn't this done a long time ago? Say when Virgil first proclaimed, "Lucky is he who has been able to understand the causes of things" (29 BC). Or when the founders of modern statistics, Francis Galton and Karl Pearson, first discovered that population data can shed light on scientific questions. There is a long tale behind their unfortunate failure to embrace causation at this juncture, which the historical sections of this book will relate. But the most serious impediment, in my opinion, has been the fundamental gap between the vocabulary in which we cast causal questions and the traditional vocabulary in which we communicate scientific theories.

To appreciate the depth of this gap, imagine the difficulties that a scientist would face in trying to express some obvious causal relationships—say, that the barometer reading B tracks the atmospheric pressure P. We can easily write down this relationship in an equation such as $B = kP$, where k is some constant of proportionality. The rules of algebra now permit us to rewrite this same equation in a wild variety of forms, for example, $P = B/k$, $k = B/P$, or $B - kP = 0$. They all mean the same thing—that if we know any two of the three quantities, the third is determined. None of the letters k, B, or P is in any mathematical way privileged over any of the others. How then can we express our strong conviction that it is the pressure that causes the barometer to change and not the other way around? And if we cannot express even this, how can we hope to express the many other causal convictions that do not have mathematical formulas, such as that the rooster's crow does

not cause the sun to rise?

My college professors could not do it and never complained. I would be willing to bet that none of yours ever did either. We now understand why: never were they shown a mathematical language of causes; nor were they shown its benefits. It is in fact an indictment of science that it has neglected to develop such a language for so many generations. Everyone knows that flipping a switch will cause a light to turn on or off and that a hot, sultry summer afternoon will cause sales to go up at the local ice-cream parlor. Why then have scientists not captured such obvious facts in formulas, as they did with the basic laws of optics, mechanics, or geometry? Why have they allowed these facts to languish in bare intuition, deprived of mathematical tools that have enabled other branches of science to flourish and mature?

Part of the answer is that scientific tools are developed to meet scientific needs. Precisely because we are so good at handling questions about switches, ice cream, and barometers, our need for special mathematical machinery to handle them was not obvious. But as scientific curiosity increased and we began posing causal questions in complex legal, business, medical, and policy-making situations, we found ourselves lacking the tools and principles that mature science should provide.

Belated awakenings of this sort are not uncommon in science. For example, until about four hundred years ago, people were quite happy with their natural ability to manage the uncertainties in daily life, from crossing a street to risking a fistfight. Only after gamblers invented intricate games of chance, sometimes carefully designed to trick us into making bad choices, did mathematicians like Blaise Pascal (1654), Pierre de Fermat (1654), and Christiaan Huygens (1657) find it necessary to develop what we today call probability theory. Likewise, only when insurance organizations demanded accurate estimates of life annuity did mathematicians like Edmond Halley (1693) and Abraham de Moivre (1725) begin looking at mortality tables to calculate life expectancies. Similarly, astronomers' demands for accurate predictions of celestial motion led Jacob Bernoulli, Pierre-Simon Laplace, and Carl Friedrich Gauss to develop a theory of errors to help us extract signals from noise. These methods were all predecessors of today's statistics.

Ironically, the need for a theory of causation began to surface

at the same time that statistics came into being. In fact, modern statistics hatched from the causal questions that Galton and Pearson asked about heredity and their ingenious attempts to answer them using cross-generational data. Unfortunately, they failed in this endeavor, and rather than pause to ask why, they declared those questions off limits and turned to developing a thriving, causality-free enterprise called statistics.

This was a critical moment in the history of science. The opportunity to equip causal questions with a language of their own came very close to being realized but was squandered. In the following years, these questions were declared unscientific and went underground. Despite heroic efforts by the geneticist Sewall Wright (1889–1988), causal vocabulary was virtually prohibited for more than half a century. And when you prohibit speech, you prohibit thought and stifle principles, methods, and tools.

Readers do not have to be scientists to witness this prohibition. In Statistics 101, every student learns to chant, "Correlation is not causation." With good reason! The rooster's crow is highly correlated with the sunrise; yet it does not cause the sunrise.

Unfortunately, statistics has fetishized this commonsense observation. It tells us that correlation is not causation, but it does not tell us what causation is. In vain will you search the index of a statistics textbook for an entry on "cause." Students are not allowed to say that X is the cause of Y—only that X and Y are "related" or "associated."

Because of this prohibition, mathematical tools to manage causal questions were deemed unnecessary, and statistics focused exclusively on how to summarize data, not on how to interpret it. A shining exception was path analysis, invented by geneticist Sewall Wright in the 1920s and a direct ancestor of the methods we will entertain in this book. However, path analysis was badly underappreciated in statistics and its satellite communities and languished for decades in its embryonic status. What should have been the first step toward causal inference remained the only step until the 1980s. The rest of statistics, including the many disciplines that looked to it for guidance, remained in the Prohibition era, falsely believing that the answers to all scientific questions reside in the data, to be unveiled through clever data-

mining tricks.

Much of this data-centric history still haunts us today. We live in an era that presumes Big Data to be the solution to all our problems. Courses in "data science" are proliferating in our universities, and jobs for "data scientists" are lucrative in the companies that participate in the "data economy." But I hope with this book to convince you that data are profoundly dumb. Data can tell you that the people who took a medicine recovered faster than those who did not take it, but they can't tell you why. Maybe those who took the medicine did so because they could afford it and would have recovered just as fast without it.

Over and over again, in science and in business, we see situations where mere data aren't enough. Most big-data enthusiasts, while somewhat aware of these limitations, continue the chase after data-centric intelligence, as if we were still in the Prohibition era.

As I mentioned earlier, things have changed dramatically in the past three decades. Nowadays, thanks to carefully crafted causal models, contemporary scientists can address problems that would have once been considered unsolvable or even beyond the pale of scientific inquiry. For example, only a hundred years ago, the question of whether cigarette smoking causes a health hazard would have been considered unscientific. The mere mention of the words "cause" or "effect" would create a storm of objections in any reputable statistical journal.

Even two decades ago, asking a statistician a question like "Was it the aspirin that stopped my headache?" would have been like asking if he believed in voodoo. To quote an esteemed colleague of mine, it would be "more of a cocktail conversation topic than a scientific inquiry." But today, epidemiologists, social scientists, computer scientists, and at least some enlightened economists and statisticians pose such questions routinely and answer them with mathematical precision. To me, this change is nothing short of a revolution. I dare to call it the Causal Revolution, a scientific shakeup that embraces rather than denies our innate cognitive gift of understanding cause and effect.

The Causal Revolution did not happen in a vacuum; it has a mathematical secret behind it which can be best described as a calculus of causation, which answers some of the hardest problems

ever asked about cause-effect relationships. I am thrilled to unveil this calculus not only because the turbulent history of its development is intriguing but even more because I expect that its full potential will be developed one day beyond what I can imagine... perhaps even by a reader of this book.

The calculus of causation consists of two languages: causal diagrams, to express what we know, and a symbolic language, resembling algebra, to express what we want to know. The causal diagrams are simply dot-and-arrow pictures that summarize our existing scientific knowledge. The dots represent quantities of interest, called "variables," and the arrows represent known or suspected causal relationships between those variables—namely, which variable "listens" to which others. These diagrams are extremely easy to draw, comprehend, and use, and the reader will find dozens of them in the pages of this book. If you can navigate using a map of one-way streets, then you can understand causal diagrams, and you can solve the type of questions posed at the beginning of this introduction.

Though causal diagrams are my tool of choice in this book, as in the last thirty-five years of my research, they are not the only kind of causal model possible. Some scientists (e.g., econometricians) like to work with mathematical equations; others (e.g., hard-core statisticians) prefer a list of assumptions that ostensibly summarizes the structure of the diagram. Regardless of language, the model should depict, however qualitatively, the process that generates the data—in other words, the cause-effect forces that operate in the environment and shape the data generated.

Side by side with this diagrammatic "language of knowledge," we also have a symbolic "language of queries" to express the questions we want answers to. For example, if we are interested in the effect of a drug (D) on lifespan (L), then our query might be written symbolically as: $P(L \mid do(D))$. In other words, what is the probability (P) that a typical patient would survive L years if made to take the drug? This question describes what epidemiologists would call an *intervention* or a *treatment* and corresponds to what we measure in a clinical trial. In many cases we may also wish to compare $P(L \mid do(D))$ with $P(L \mid do(\text{not-}D))$; the latter describes patients denied treatment, also called the "control" patients. The

do-operator signifies that we are dealing with an intervention rather than a passive observation; classical statistics has nothing remotely similar to this operator.

We must invoke an intervention operator $do(D)$ to ensure that the observed change in Lifespan L is due to the drug itself and is not confounded with other factors that tend to shorten or lengthen life. If, instead of intervening, we let the patient himself decide whether to take the drug, those other factors might influence his decision, and lifespan differences between taking and not taking the drug would no longer be solely due to the drug. For example, suppose only those who were terminally ill took the drug. Such persons would surely differ from those who did not take the drug, and a comparison of the two groups would reflect differences in the severity of their disease rather than the effect of the drug. By contrast, forcing patients to take or refrain from taking the drug, regardless of preconditions, would wash away preexisting differences and provide a valid comparison.

Mathematically, we write the observed frequency of Lifespan L among patients who voluntarily take the drug as $P(L \mid D)$, which is the standard conditional probability used in statistical textbooks. This expression stands for the probability (P) of Lifespan L conditional on seeing the patient take Drug D. Note that $P(L \mid D)$ may be totally different from $P(L \mid do(D))$. This difference between seeing and doing is fundamental and explains why we do not regard the falling barometer to be a cause of the coming storm. Seeing the barometer fall increases the probability of the storm, while forcing it to fall does not affect this probability.

This confusion between seeing and doing has resulted in a fountain of paradoxes, some of which we will entertain in this book. A world devoid of $P(L \mid do(D))$ and governed solely by $P(L \mid D)$ would be a strange one indeed. For example, patients would avoid going to the doctor to reduce the probability of being seriously ill; cities would dismiss their firefighters to reduce the incidence of fires; doctors would recommend a drug to male and female patients but not to patients with undisclosed gender; and so on. It is hard to believe that less than three decades ago science did operate in such a world: the *do*-operator did not exist.

One of the crowning achievements of the Causal Revolution has been to explain how to predict the effects of an intervention

without actually enacting it. It would never have been possible if we had not, first of all, defined the *do*-operator so that we can ask the right question and, second, devised a way to emulate it by noninvasive means.

When the scientific question of interest involves retrospective thinking, we call on another type of expression unique to causal reasoning called a counterfactual. For example, suppose that Joe took Drug *D* and died a month later; our question of interest is whether the drug might have caused his death. To answer this question, we need to imagine a scenario in which Joe was about to take the drug but changed his mind. Would he have lived?

Again, classical statistics only summarizes data, so it does not provide even a language for asking that question. Causal inference provides a notation and, more importantly, offers a solution. As with predicting the effect of interventions (mentioned above), in many cases we can emulate human retrospective thinking with an algorithm that takes what we know about the observed world and produces an answer about the counterfactual world. This "algorithmization of counterfactuals" is another gem uncovered by the Causal Revolution.

Counterfactual reasoning, which deals with what-ifs, might strike some readers as unscientific. Indeed, empirical observation can never confirm or refute the answers to such questions. Yet our minds make very reliable and reproducible judgments all the time about what might be or might have been. We all understand, for instance, that had the rooster been silent this morning, the sun would have risen just as well. This consensus stems from the fact that counterfactuals are not products of whimsy but reflect the very structure of our world model. Two people who share the same causal model will also share all counterfactual judgments.

Counterfactuals are the building blocks of moral behavior as well as scientific thought. The ability to reflect on one's past actions and envision alternative scenarios is the basis of free will and social responsibility. The algorithmization of counterfactuals invites thinking machines to benefit from this ability and participate in this (until now) uniquely human way of thinking about the world.

My mention of thinking machines in the last paragraph is intentional. I came to this subject as a computer scientist working

in the area of artificial intelligence, which entails two points of departure from most of my colleagues in the causal inference arena. First, in the world of AI, you do not really understand a topic until you can teach it to a mechanical robot. That is why you will find me emphasizing and reemphasizing notation, language, vocabulary, and grammar. For example, I obsess over whether we can express a certain claim in a given language and whether one claim follows from others. It is amazing how much one can learn from just following the grammar of scientific utterances. My emphasis on language also comes from a deep conviction that language shapes our thoughts. You cannot answer a question that you cannot ask, and you cannot ask a question that you have no words for. As a student of philosophy and computer science, my attraction to causal inference has largely been triggered by the excitement of seeing an orphaned scientific language making it from birth to maturity.

My background in machine learning has given me yet another incentive for studying causation. In the late 1980s, I realized that machines' lack of understanding of causal relations was perhaps the biggest roadblock to giving them human-level intelligence. In the last chapter of this book, I will return to my roots, and together we will explore the implications of the Causal Revolution for artificial intelligence. I believe that strong AI is an achievable goal and one not to be feared precisely because causality is part of the solution. A causal reasoning module will give machines the ability to reflect on their mistakes, to pinpoint weaknesses in their software, to function as moral entities, and to converse naturally with humans about their own choices and intentions.

A BLUEPRINT OF REALITY

In our era, readers have no doubt heard terms like "knowledge," "information," "intelligence," and "data," and some may feel confused about the differences between them or how they interact. Now I am proposing to throw another term, "causal model," into the mix, and the reader may justifiably wonder if this will only add to the confusion.

It will not! In fact, it will anchor the elusive notions of science,

knowledge, and data in a concrete and meaningful setting, and will enable us to see how the three work together to produce answers to difficult scientific questions. Figure I.1 shows a blueprint for a "causal inference engine" that might handle causal reasoning for a future artificial intelligence. It's important to realize that this is not only a blueprint for the future but also a guide to how causal models work in scientific applications today and how they interact with data.

The inference engine is a machine that accepts three different kinds of inputs—Assumptions, Queries, and Data—and produces three kinds of outputs. The first of the outputs is a Yes/No decision as to whether the given query can in theory be answered under the existing causal model, assuming perfect and unlimited data. If the answer is Yes, the inference engine next produces an Estimand. This is a mathematical formula that can be thought of as a recipe for generating the answer from any hypothetical data, whenever they are available. Finally, after the inference engine has received the Data input, it will use the recipe to produce an actual Estimate for the answer, along with statistical estimates of the amount of uncertainty in that estimate. This uncertainty reflects the limited size of the data set as well as possible measurement errors or missing data.

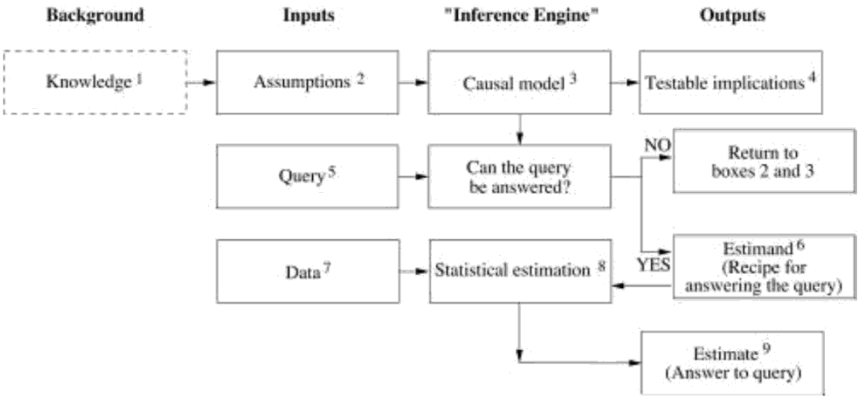


FIGURE I. How an "inference engine" combines data with causal knowledge to produce answers to queries of interest. The dashed box is not part of the engine but is required for building it. Arrows could also be drawn from boxes 4 and 9 to box 1, but I have opted to keep the diagram simple.

To dig more deeply into the chart, I have labeled the boxes 1

through 9, which I will annotate in the context of the query "What is the effect of Drug *D* on Lifespan *L*?"

1. "Knowledge" stands for traces of experience the reasoning agent has had in the past, including past observations, past actions, education, and cultural mores, that are deemed relevant to the query of interest. The dotted box around "Knowledge" indicates that it remains implicit in the mind of the agent and is not explicated formally in the model.
2. Scientific research always requires simplifying assumptions, that is, statements which the researcher deems worthy of making explicit on the basis of the available Knowledge. While most of the researcher's knowledge remains implicit in his or her brain, only Assumptions see the light of day and are encapsulated in the model. They can in fact be read from the model, which has led some logicians to conclude that a model is nothing more than a list of assumptions. Computer scientists take exception to this claim, noting that how assumptions are represented can make a profound difference in one's ability to specify them correctly, draw conclusions from them, and even extend or modify them in light of compelling evidence.
3. Various options exist for causal models: causal diagrams, structural equations, logical statements, and so forth. I am strongly sold on causal diagrams for nearly all applications, primarily due to their transparency but also due to the explicit answers they provide to many of the questions we wish to ask. For the purpose of constructing the diagram, the definition of "causation" is simple, if a little metaphorical: a variable X is a cause of Y if Y "listens" to X and determines its value in response to what it hears. For example, if we suspect that a patient's Lifespan L "listens" to whether Drug D was

taken, then we call D a cause of L and draw an arrow from D to L in a causal diagram. Naturally, the answer to our query about D and L is likely to depend on other variables as well, which must also be represented in the diagram along with their causes and effects. (Here, we will denote them collectively by Z.)

4. The listening pattern prescribed by the paths of the causal model usually results in observable patterns or dependencies in the data. These patterns are called "testable implications" because they can be used for testing the model. These are statements like "There is no path connecting D and L," which translates to a statistical statement, "D and L are independent," that is, finding D does not change the likelihood of L. If the data contradict this implication, then we need to revise our model. Such revisions require another engine, which obtains its inputs from boxes 4 and 7 and computes the "degree of fitness," that is, the degree to which the Data are compatible with the model's assumptions. For simplicity, I did not show this second engine in Figure I.1.
5. Queries submitted to the inference engine are the scientific questions that we want to answer. They must be formulated in causal vocabulary. For example, what is $P(L \mid do(D))$? One of the main accomplishments of the Causal Revolution has been to make this language scientifically transparent as well as mathematically rigorous.
6. "Estimand" comes from Latin, meaning "that which is to be estimated." This is a statistical quantity to be estimated from the data that, once estimated, can legitimately represent the answer to our query. While written as a probability formula—for example, $P(L \mid D, Z) \times P(Z)$—it is in fact a recipe for answering the causal query from the type of data

we have, once it has been certified by the engine.

It's very important to realize that, contrary to traditional estimation in statistics, some queries may not be answerable under the current causal model, even after the collection of any amount of data. For example, if our model shows that both D and L depend on a third variable Z (say, the stage of a disease), and if we do not have any way to measure Z, then the query $P(L \mid do(D))$ cannot be answered. In that case it is a waste of time to collect data. Instead we need to go back and refine the model, either by adding new scientific knowledge that might allow us to estimate Z or by making simplifying assumptions (at the risk of being wrong)—for example, that the effect of Z on D is negligible.

7. Data are the ingredients that go into the estimand recipe. It is critical to realize that data are profoundly dumb about causal relationships. They tell us about quantities like $P(L \mid D)$ or $P(L \mid D, Z)$. It is the job of the estimand to tell us how to bake these statistical quantities into one expression that, based on the model assumptions, is logically equivalent to the causal query—say, $P(L \mid do(D))$.

   Notice that the whole notion of estimands and in fact the whole top part of Figure I does not exist in traditional methods of statistical analysis. There, the estimand and the query coincide. For example, if we are interested in the proportion of people among those with Lifespan L who took the Drug D, we simply write this query as $P(D \mid L)$. The same quantity would be our estimand. This already specifies what proportions in the data need to be estimated and requires no causal knowledge. For this reason, some statisticians to this day find it extremely hard to understand why some knowledge lies outside the province of statistics and why data alone cannot make up for lack of scientific knowledge.

8. The estimate is what comes out of the oven. However, it is only approximate because of one other real-world fact about data: they are always only a finite sample from a theoretically infinite population. In our running example, the sample consists of the patients we choose to study. Even if we choose them at random, there is always some chance that the proportions measured in the sample are not representative of the proportions in the population at large. Fortunately, the discipline of statistics, nowadays empowered by advanced techniques of machine learning, gives us many, many ways to manage this uncertainty—parametric and semi-parametric models, maximum likelihood methods, and propensity scores, are often used to smooth the sparse data.
9. In the end, if our model is correct and our data are sufficient, we get an answer to our causal query, such as "Drug *D* increases the Lifespan *L* of diabetic Patients *Z* by 30 percent, plus or minus 20 percent." Hooray! The answer will also add to our scientific knowledge (box 1) and, if things did not go the way we expected, might suggest some improvements to our causal model (box 3).

This flowchart may look complicated at first, and you might wonder whether it is really necessary. Indeed, in our ordinary lives, we are somehow able to make causal judgments without consciously going through such a complicated process and certainly without resorting to the mathematics of probabilities and proportions. Our causal intuition alone is usually sufficient for handling the kind of uncertainty we find in household routines or even in our professional lives. But if we want to teach a dumb robot to think causally, or if we are pushing the frontiers of scientific knowledge, where we do not have intuition to guide us, then a carefully structured procedure like this is mandatory.

I especially want to highlight the role of data in the above process. First, notice that we collect data only after we posit the

causal model, after we state the scientific query we wish to answer, and after we derive the estimand. This contrasts with the traditional statistical approach, mentioned above, which does not even have a causal model.

But our present-day scientific world presents a new challenge to sound reasoning about causes and effects. While awareness of the need for a causal model has grown by leaps and bounds among the sciences, many researchers in artificial intelligence would like to skip the hard step of constructing or acquiring a causal model and rely solely on data for all cognitive tasks. The hope—and at present, it is usually a silent one—is that the data themselves will guide us to the right answers whenever causal questions come up.

I am an outspoken skeptic of this trend because I know how profoundly dumb data are about causes and effects. For example, information about the effects of actions or interventions is simply not available in raw data, unless it is collected by controlled experimental manipulation. By contrast, if we are in possession of a causal model, we can often predict the result of an intervention from hands-off, intervention-free data.

The case for causal models becomes even more compelling when we seek to answer counterfactual queries such as "What would have happened had we acted differently?" We will discuss counterfactuals in great detail because they are the most challenging queries for any artificial intelligence. They are also at the core of the cognitive advances that made us human and the imaginative abilities that have made science possible. We will also explain why any query about the mechanism by which causes transmit their effects—the most prototypical "Why?" question—is actually a counterfactual question in disguise. Thus, if we ever want robots to answer "Why?" questions or even understand what they mean, we must equip them with a causal model and teach them how to answer counterfactual queries, as in Figure I.1.

Another advantage causal models have that data mining and deep learning lack is adaptability. Note that in Figure I.1, the estimand is computed on the basis of the causal model alone, prior to an examination of the specifics of the data. This makes the causal inference engine supremely adaptable, because the estimand computed is good for any data that are compatible with the qualitative model, regardless of the numerical relationships

among the variables.

To see why this adaptability is important, compare this engine with a learning agent—in this instance a human, but in other cases perhaps a deep-learning algorithm or maybe a human using a deep-learning algorithm—trying to learn solely from the data. By observing the outcome L of many patients given Drug D, she is able to predict the probability that a patient with characteristics Z will survive L years. Now she is transferred to a different hospital, in a different part of town, where the population characteristics (diet, hygiene, work habits) are different. Even if these new characteristics merely modify the numerical relationships among the variables recorded, she will still have to retrain herself and learn a new prediction function all over again. That's all that a deep-learning program can do: fit a function to data. On the other hand, if she possessed a model of how the drug operated and its causal structure remained intact in the new location, then the estimand she obtained in training would remain valid. It could be applied to the new data to generate a new population-specific prediction function.

Many scientific questions look different "through a causal lens," and I have delighted in playing with this lens, which over the last twenty-five years has been increasingly empowered by new insights and new tools. I hope and believe that readers of this book will share in my delight. Therefore, I'd like to close this introduction with a preview of some of the coming attractions in this book.

Chapter 1 assembles the three steps of observation, intervention, and counterfactuals into the Ladder of Causation, the central metaphor of this book. It will also expose you to the basics of reasoning with causal diagrams, our main modeling tool, and set you well on your way to becoming a proficient causal reasoner—in fact, you will be far ahead of generations of data scientists who attempted to interpret data through a model-blind lens, oblivious to the distinctions that the Ladder of Causation illuminates.

Chapter 2 tells the bizarre story of how the discipline of statistics inflicted causal blindness on itself, with far-reaching effects for all sciences that depend on data. It also tells the story of one of the great heroes of this book, the geneticist Sewall Wright, who in the 1920s drew the first causal diagrams and for many

years was one of the few scientists who dared to take causality seriously.

Chapter 3 relates the equally curious story of how I became a convert to causality through my work in AI and particularly on Bayesian networks. These were the first tool that allowed computers to think in "shades of gray"—and for a time I believed they held the key to unlocking AI. Toward the end of the 1980s I became convinced that I was wrong, and this chapter tells of my journey from prophet to apostate. Nevertheless, Bayesian networks remain a very important tool for AI and still encapsulate much of the mathematical foundation of causal diagrams. In addition to a gentle, causality-minded introduction to Bayes's rule and Bayesian methods of reasoning, Chapter 3 will entertain the reader with examples of real-life applications of Bayesian networks.

Chapter 4 tells about the major contribution of statistics to causal inference: the randomized controlled trial (RCT). From a causal perspective, the RCT is a man-made tool for uncovering the query $P(L \mid do(D))$, which is a property of nature. Its main purpose is to disassociate variables of interest (say, D and L) from other variables (Z) that would otherwise affect them both. Disarming the distortions, or "confounding," produced by such lurking variables has been a century-old problem. This chapter walks the reader through a surprisingly simple solution to the general confounding problem, which you will grasp in ten minutes of playfully tracing paths in a diagram.

Chapter 5 gives an account of a seminal moment in the history of causation and indeed the history of science, when statisticians struggled with the question of whether smoking causes lung cancer. Unable to use their favorite tool, the randomized controlled trial, they struggled to agree on an answer or even on how to make sense of the question. The smoking debate brings the importance of causality into its sharpest focus. Millions of lives were lost or shortened because scientists did not have an adequate language or methodology for answering causal questions.

Chapter 6 will, I hope, be a welcome diversion for the reader after the serious matters of Chapter 5. This is a chapter of paradoxes: the Monty Hall paradox, Simpson's paradox, Berkson's paradox, and others. Classical paradoxes like these can be enjoyed

as brainteasers, but they have a serious side too, especially when viewed from a causal perspective. In fact, almost all of them represent clashes with causal intuition and therefore reveal the anatomy of that intuition. They were canaries in the coal mine that should have alerted scientists to the fact that human intuition is grounded in causal, not statistical, logic. I believe that the reader will enjoy this new twist on his or her favorite old paradoxes.

Chapters 7 to 9 finally take readers on a thrilling ascent of the Ladder of Causation. We start in Chapter 7 with questions about intervention and explain how my students and I went through a twenty-year struggle to automate the answers to *do*-type questions. We succeeded, and this chapter explains the guts of the "causal inference engine," which produces the yes/no answer and the estimand in Figure I.1. Studying this engine will empower the reader to spot certain patterns in the causal diagram that deliver immediate answers to the causal query. These patterns are called back-door adjustment, front-door adjustment, and instrumental variables, the workhorses of causal inference in practice.

Chapter 8 takes you to the top of the ladder by discussing counterfactuals. These have been seen as a fundamental part of causality at least since 1748, when Scottish philosopher David Hume proposed the following somewhat contorted definition of causation: "We may define a cause to be an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed." David Lewis, a philosopher at Princeton University who died in 2001, pointed out that Hume really gave two definitions, not one, the first of regularity (i.e., the cause is regularly followed by the effect) and the second of the counterfactual ("if the first object had not been..."). While philosophers and scientists had mostly paid attention to the regularity definition, Lewis argued that the counterfactual definition aligns more closely with human intuition: "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it."

Readers will be excited to find out that we can now move past the academic debates and compute an actual value (or probability) for any counterfactual query, no matter how convoluted. Of special

interest are questions concerning necessary and sufficient causes of observed events. For example, how likely is it that the defendant's action was a necessary cause of the claimant's injury? How likely is it that man-made climate change is a sufficient cause of a heat wave?

Finally, Chapter 9 discusses the topic of mediation. You may have wondered, when we talked about drawing arrows in a causal diagram, whether we should draw an arrow from Drug D to Lifespan L if the drug affects lifespan only by way of its effect on blood pressure Z (a mediator). In other words, is the effect of D on L direct or indirect? And if both, how do we assess their relative importance? Such questions are not only of great scientific interest but also have practical ramifications; if we understand the mechanism through which a drug acts, we might be able to develop other drugs with the same effect that are cheaper or have fewer side effects. The reader will be pleased to discover how this age-old quest for a mediation mechanism has been reduced to an algebraic exercise and how scientists are using the new tools in the causal tool kit to solve such problems.

Chapter 10 brings the book to a close by coming back to the problem that initially led me to causation: the problem of automating human-level intelligence (sometimes called "strong AI"). I believe that causal reasoning is essential for machines to communicate with us in our own language about policies, experiments, explanations, theories, regret, responsibility, free will, and obligations—and, eventually, to make their own moral decisions.

If I could sum up the message of this book in one pithy phrase, it would be that you are smarter than your data. Data do not understand causes and effects; humans do. I hope that the new science of causal inference will enable us to better understand how we do it, because there is no better way to understand ourselves than by emulating ourselves. In the age of computers, this new understanding also brings with it the prospect of amplifying our innate abilities so that we can make better sense of data, be it big or small.

1

THE LADDER OF CAUSATION

IN the beginning...

I was probably six or seven years old when I first read the story of Adam and Eve in the Garden of Eden. My classmates and I were not at all surprised by God's capricious demands, forbidding them to eat from the Tree of Knowledge. Deities have their reasons, we thought. What we were more intrigued by was the idea that as soon as they ate from the Tree of Knowledge, Adam and Eve became conscious, like us, of their nakedness.

As teenagers, our interest shifted slowly to the more philosophical aspects of the story. (Israeli students read Genesis several times a year.) Of primary concern to us was the notion that the emergence of human knowledge was not a joyful process but a painful one, accompanied by disobedience, guilt, and punishment. Was it worth giving up the carefree life of Eden? some asked. Were the agricultural and scientific revolutions that followed worth the economic hardships, wars, and social injustices that modern life entails?

Don't get me wrong: we were no creationists; even our teachers were Darwinists at heart. We knew, however, that the author who choreographed the story of Genesis struggled to answer the most pressing philosophical questions of his time. We likewise suspected that this story bore the cultural footprints of the actual process by which *Homo sapiens* gained dominion over our planet. What, then, was the sequence of steps in this speedy,

super-evolutionary process?

My interest in these questions waned in my early career as a professor of engineering but was reignited suddenly in the 1990s, when, while writing my book *Causality*, I confronted the Ladder of Causation.

As I reread Genesis for the hundredth time, I noticed a nuance that had somehow eluded my attention for all those years. When God finds Adam hiding in the garden, he asks, "Have you eaten from the tree which I forbade you?" And Adam answers, "The woman you gave me for a companion, she gave me fruit from the tree and I ate." "What is this you have done?" God asks Eve. She replies, "The serpent deceived me, and I ate."

As we know, this blame game did not work very well on the Almighty, who banished both of them from the garden. But here is the point I had missed before: God asked "what," and they answered "why." God asked for the facts, and they replied with explanations. Moreover, both were thoroughly convinced that naming causes would somehow paint their actions in a different light. Where did they get this idea?

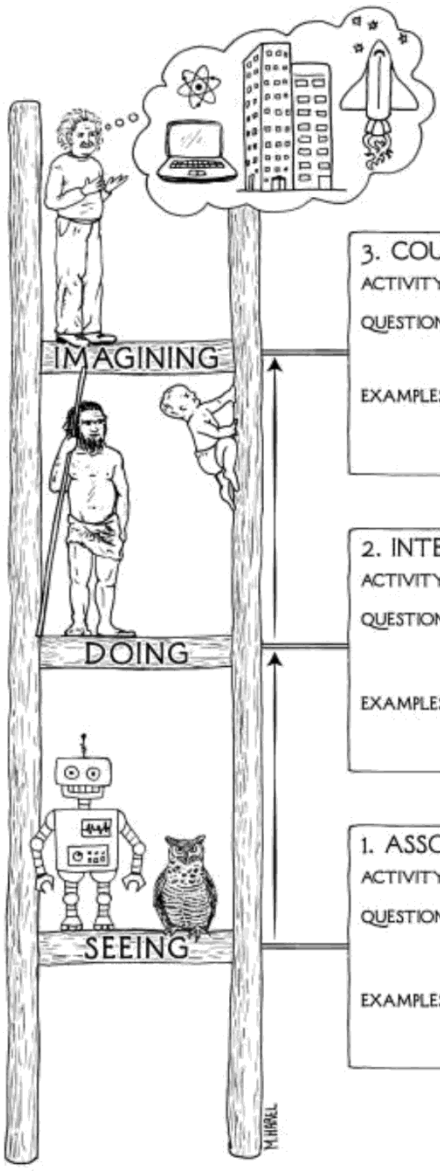
For me, these nuances carried three profound implications. First, very early in our evolution, we humans realized that the world is not made up only of dry facts (what we might call data today); rather, these facts are glued together by an intricate web of cause-effect relationships. Second, causal explanations, not dry facts, make up the bulk of our knowledge, and should be the cornerstone of machine intelligence. Finally, our transition from processors of data to makers of explanations was not gradual; it was a leap that required an external push from an uncommon fruit. This matched perfectly with what I had observed theoretically in the Ladder of Causation: No machine can derive explanations from raw data. It needs a push.

If we seek confirmation of these messages from evolutionary science, we won't find the Tree of Knowledge, of course, but we still see a major unexplained transition. We understand now that humans evolved from apelike ancestors over a period of 5 million to 6 million years and that such gradual evolutionary processes are not uncommon to life on earth. But in roughly the last 50,000 years, something unique happened, which some call the Cognitive Revolution and others (with a touch of irony) call the Great Leap

distinct levels of cognitive ability: seeing, doing, and imagining.

The first, seeing or observing, entails detection of regularities in our environment and is shared by many animals as well as early humans before the Cognitive Revolution. The second, doing, entails predicting the effect(s) of deliberate alterations of the environment and choosing among these alterations to produce a desired outcome. Only a small handful of species have demonstrated elements of this skill. Use of tools, provided it is intentional and not just accidental or copied from ancestors, could be taken as a sign of reaching this second level. Yet even tool users do not necessarily possess a "theory" of their tool that tells them why it works and what to do when it doesn't. For that, you need to have achieved a level of understanding that permits imagining. It was primarily this third level that prepared us for further revolutions in agriculture and science and led to a sudden and drastic change in our species' impact on the planet.

I cannot prove this, but I can prove mathematically that the three levels differ fundamentally, each unleashing capabilities that the ones below it do not. The framework I use to show this goes back to Alan Turing, the pioneer of research in artificial intelligence (AI), who proposed to classify a cognitive system in terms of the queries it can answer. This approach is exceptionally fruitful when we are talking about causality because it bypasses long and unproductive discussions of what exactly causality is and focuses instead on the concrete and answerable question "What can a causal reasoner do?" Or more precisely, what can an organism possessing a causal model compute that one lacking such a model cannot?



3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

FIGURE 1.2. The Ladder of Causation, with representative organisms at each level. Most animals, as well as present-day learning machines, are on the first rung, learning from association. Tool users, such as early humans, are on the second rung if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumably this is how babies acquire much of their causal knowledge. Counterfactual learners, on the top rung, can imagine worlds that do not exist and infer reasons for observed phenomena. (*Source:* Drawing by Maayan Harel.)

While Turing was looking for a binary classification—human or nonhuman—ours has three tiers, corresponding to progressively more powerful causal queries. Using these criteria, we can assemble the three levels of queries into one Ladder of Causation (Figure 1.2), a metaphor that we will return to again and again.

Let's take some time to consider each rung of the ladder in detail. At the first level, association, we are looking for regularities in observations. This is what an owl does when observing how a rat moves and figuring out where the rodent is likely to be a moment later, and it is what a computer Go program does when it studies a database of millions of Go games so that it can figure out which moves are associated with a higher percentage of wins. We say that one event is associated with another if observing one changes the likelihood of observing the other.

The first rung of the ladder calls for predictions based on passive observations. It is characterized by the question "What if I see ...?" For instance, imagine a marketing director at a department store who asks, "How likely is a customer who bought toothpaste to also buy dental floss?" Such questions are the bread and butter of statistics, and they are answered, first and foremost, by collecting and analyzing data. In our case, the question can be answered by first taking the data consisting of the shopping behavior of all customers, selecting only those who bought toothpaste, and, focusing on the latter group, computing the proportion who also bought dental floss. This proportion, also known as a "conditional probability," measures (for large data) the degree of association between "buying toothpaste" and "buying floss." Symbolically, we can write it as $P(floss \mid toothpaste)$. The "P" stands for "probability," and the vertical line means "given that you see."

Statisticians have developed many elaborate methods to reduce a large body of data and identify associations between variables. "Correlation" or "regression," a typical measure of association mentioned often in this book, involves fitting a line to a collection of data points and taking the slope of that line. Some associations might have obvious causal interpretations; others may not. But statistics alone cannot tell which is the cause and which is the effect, toothpaste or floss. From the point of view of the sales manager, it may not really matter. Good predictions need

not have good explanations. The owl can be a good hunter without understanding why the rat always goes from point *A* to point *B*.

Some readers may be surprised to see that I have placed present-day learning machines squarely on rung one of the Ladder of Causation, sharing the wisdom of an owl. We hear almost every day, it seems, about rapid advances in machine learning systems—self-driving cars, speech-recognition systems, and, especially in recent years, deep-learning algorithms (or deep neural networks). How could they still be only at level one?

The successes of deep learning have been truly remarkable and have caught many of us by surprise. Nevertheless, deep learning has succeeded primarily by showing that certain questions or tasks we thought were difficult are in fact not. It has not addressed the truly difficult questions that continue to prevent us from achieving humanlike AI. As a result the public believes that "strong AI," machines that think like humans, is just around the corner or maybe even here already. In reality, nothing could be farther from the truth. I fully agree with Gary Marcus, a neuroscientist at New York University, who recently wrote in the *New York Times* that the field of artificial intelligence is "bursting with microdiscoveries"—the sort of things that make good press releases—but machines are still disappointingly far from humanlike cognition. My colleague in computer science at the University of California, Los Angeles, Adnan Darwiche, has titled a position paper "Human-Level Intelligence or Animal-Like Abilities?" which I think frames the question in just the right way. The goal of strong AI is to produce machines with humanlike intelligence, able to converse with and guide humans. Deep learning has instead given us machines with truly impressive abilities but no intelligence. The difference is profound and lies in the absence of a model of reality.

Just as they did thirty years ago, machine learning programs (including those with deep neural networks) operate almost entirely in an associational mode. They are driven by a stream of observations to which they attempt to fit a function, in much the same way that a statistician tries to fit a line to a collection of points. Deep neural networks have added many more layers to the complexity of the fitted function, but raw data still drives the fitting process. They continue to improve in accuracy as more data

respond by changing from "headache" to "no headache."

While reasoning about interventions is an important step on the causal ladder, it still does not answer all questions of interest. We might wonder, My headache is gone now, but why? Was it the aspirin I took? The food I ate? The good news I heard? These queries take us to the top rung of the Ladder of Causation, the level of counterfactuals, because to answer them we must go back in time, change history, and ask, "What would have happened if I had not taken the aspirin?" No experiment in the world can deny treatment to an already treated person and compare the two outcomes, so we must import a whole new kind of knowledge.

Counterfactuals have a particularly problematic relationship with data because data are, by definition, facts. They cannot tell us what will happen in a counterfactual or imaginary world where some observed facts are bluntly negated. Yet the human mind makes such explanation-seeking inferences reliably and repeatably. Eve did it when she identified "The serpent deceived me" as the reason for her action. This ability most distinguishes human from animal intelligence, as well as from model-blind versions of AI and machine learning.

You may be skeptical that science can make any useful statement about "would haves," worlds that do not exist and things that have not happened. But it does and always has. The laws of physics, for example, can be interpreted as counterfactual assertions, such as "Had the weight on this spring doubled, its length would have doubled as well" (Hooke's law). This statement is, of course, backed by a wealth of experimental (rung-two) evidence, derived from hundreds of springs, in dozens of laboratories, on thousands of different occasions. However, once anointed as a "law," physicists interpret it as a functional relationship that governs this very spring, at this very moment, under hypothetical values of the weight. All of these different worlds, where the weight is x pounds and the length of the spring is L_X inches, are treated as objectively knowable and simultaneously active, even though only one of them actually exists.

Going back to the toothpaste example, a top-rung question would be "What is the probability that a customer who bought toothpaste would still have bought it if we had doubled the price?"

We are comparing the real world (where we know that the customer bought the toothpaste at the current price) to a fictitious world (where the price is twice as high).

The rewards of having a causal model that can answer counterfactual questions are immense. Finding out why a blunder occurred allows us to take the right corrective measures in the future. Finding out why a treatment worked on some people and not on others can lead to a new cure for a disease. Answering the question "What if things had been different?" allows us to learn from history and the experience of others, something that no other species appears to do. It is not surprising that the ancient Greek philosopher Democritus (460–370 BC) said, "I would rather discover one cause than be the King of Persia."

The position of counterfactuals at the top of the Ladder of Causation explains why I place such emphasis on them as a key moment in the evolution of human consciousness. I totally agree with Yuval Harari that the depiction of imaginary creatures was a manifestation of a new ability, which he calls the Cognitive Revolution. His prototypical example is the Lion Man sculpture, found in Stadel Cave in southwestern Germany and now held at the Ulm Museum (see Figure 1.3). The Lion Man, roughly 40,000 years old, is a mammoth tusk sculpted into the form of a chimera, half man and half lion.

We do not know who sculpted the Lion Man or what its purpose was, but we do know that anatomically modern humans made it and that it represents a break with any art or craft that had gone before. Previously, humans had fashioned tools and representational art, from beads to flutes to spear points to elegant carvings of horses and other animals. The Lion Man is different: a creature of pure imagination.

FIGURE 1.3. The Lion Man of Stadel Cave. The earliest known representation of an imaginary creature (half man and half lion), it is emblematic of a newly developed cognitive ability, the capacity to reason about counterfactuals. (*Source:* Photo by Yvonne Mühleis, courtesy of State Office for Cultural Heritage Baden-Württemberg/Ulmer Museum, Ulm, Germany.)

As a manifestation of our newfound ability to imagine things that have never existed, the Lion Man is the precursor of every philosophical theory, scientific discovery, and technological innovation, from microscopes to airplanes to computers. Every one of these had to take shape in someone's imagination before it was realized in the physical world.

This leap forward in cognitive ability was as profound and important to our species as any of the anatomical changes that made us human. Within 10,000 years after the Lion Man's creation, all other hominids (except for the very geographically isolated Flores hominids) had become extinct. And humans have continued to change the natural world with incredible speed, using our imagination to survive, adapt, and ultimately take over. The advantage we gained from imagining counterfactuals was the same then as it is today: flexibility, the ability to reflect and improve on past actions, and, perhaps even more significant, our willingness to take responsibility for past and current actions.

As shown in Figure 1.2, the characteristic queries for the third rung of the Ladder of Causation are "What if I had done...?" and "Why?" Both involve comparing the observed world to a counterfactual world. Experiments alone cannot answer such questions. While rung one deals with the seen world, and rung two deals with a brave new world that is seeable, rung three deals with a world that cannot be seen (because it contradicts what is seen). To bridge the gap, we need a model of the underlying causal process, sometimes called a "theory" or even (in cases where we

insights into how the knowledge ought to be acquired, be it from data or a programmer.

When I describe the mini-Turing test, people commonly claim that it can easily be defeated by cheating. For example, take the list of all possible questions, store their correct answers, and then read them out from memory when asked. There is no way to distinguish (so the argument goes) between a machine that stores a dumb question-answer list and one that answers the way that you and I do—that is, by understanding the question and producing an answer using a mental causal model. So what would the mini-Turing test prove, if cheating is so easy?

The philosopher John Searle introduced this cheating possibility, known as the "Chinese Room" argument, in 1980 to challenge Turing's claim that the ability to fake intelligence amounts to having intelligence. Searle's challenge has only one flaw: cheating is not easy; in fact, it is impossible. Even with a small number of variables, the number of possible questions grows astronomically. Say that we have ten causal variables, each of which takes only two values (0 or 1). We could ask roughly 30 million possible queries, such as "What is the probability that the outcome is 1, given that we *see* variable X equals 1 and we *make* variable Y equal 0 and variable Z equal 1?" If there were more variables, or more than two states for each one, the number of possibilities would grow beyond our ability to even imagine. Searle's list would need more entries than the number of atoms in the universe. So, clearly a dumb list of questions and answers can never simulate the intelligence of a child, let alone an adult.

Humans must have some compact representation of the information needed in their brains, as well as an effective procedure to interpret each question properly and extract the right answer from the stored representation. To pass the mini-Turing test, therefore, we need to equip machines with a similarly efficient representation and answer-extraction algorithm.

Such a representation not only exists but has childlike simplicity: a causal diagram. We have already seen one example, the diagram for the mammoth hunt. Considering the extreme ease with which people can communicate their knowledge with dot-and-arrow diagrams, I believe that our brains indeed use a representation like this. But more importantly for our purposes,

these models pass the mini-Turing test; no other model is known to do so. Let's look at some examples.

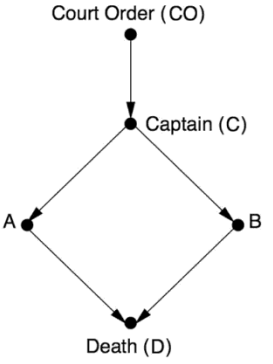


FIGURE 1.4 Causal diagram for the firing squad example. *A* and *B* represent (the actions of) Soldiers *A* and *B*.

Suppose that a prisoner is about to be executed by a firing squad. A certain chain of events must occur for this to happen. First, the court orders the execution. The order goes to a captain, who signals the soldiers on the firing squad (*A* and *B*) to fire. We'll assume that they are obedient and expert marksmen, so they only fire on command, and if either one of them shoots, the prisoner dies.

Figure 1.4 shows a diagram representing the story I just told. Each of the unknowns (*CO*, *C*, *A*, *B*, *D*) is a true/false variable. For example, *D* = true means the prisoner is dead; *D* = false means the prisoner is alive. *CO* = false means the court order was not issued; *CO* = true means it was, and so on.

Using this diagram, we can start answering causal questions from different rungs of the ladder. First, we can answer questions of association (i.e., what one fact tells us about another). If the prisoner is dead, does that mean the court order was given? We (or a computer) can inspect the graph, trace the rules behind each of the arrows, and, using standard logic, conclude that the two soldiers wouldn't have fired without the captain's command. Likewise, the captain wouldn't have given the command if he didn't have the order in his possession. Therefore the answer to our query is yes. Alternatively, suppose we find out that *A* fired. What does that tell us about *B*? By following the arrows, the computer concludes that *B* must have fired too. (*A* would not have

fired if the captain hadn't signaled, so *B* must have fired as well.) This is true even though *A* does not cause *B* (there is no arrow from *A* to *B*).

Going up the Ladder of Causation, we can ask questions about intervention. What if Soldier *A* decides on his own initiative to fire, without waiting for the captain's command? Will the prisoner be dead or alive? This question in fact already has a contradictory flavor to it. I just told you that *A* only shoots if commanded to, and yet now we are asking what happens if he fired without a command. If you're just using the rules of logic, as computers typically do, the question is meaningless. As the robot in the 1960s sci-fi TV series *Lost in Space* used to say in such situations, "That does not compute."

If we want our computer to understand causation, we have to teach it how to break the rules. We have to teach it the difference between merely observing an event and making it happen. "Whenever you make an event happen," we tell the computer, "remove all arrows that point to that event and continue the analysis by ordinary logic, as if the arrows had never been there." Thus, we erase all the arrows leading into the intervened variable (*A*). We also set that variable manually to its prescribed value (true). The rationale for this peculiar "surgery" is simple: making an event happen means that you emancipate it from all other influences and subject it to one and only one influence—that which enforces its happening.

Figure 1.5 shows the causal diagram that results from our example. This intervention leads inevitably to the prisoner's death. That is the causal function behind the arrow leading from *A* to *D*.

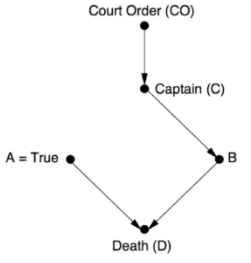


FIGURE 1.5. Reasoning about interventions. Soldier *A* decides to fire; arrow from *C* to *A* is deleted, and *A* is assigned the value true.

Note that this conclusion agrees with our intuitive judgment that *A*'s unauthorized firing will lead to the prisoner's death, because the surgery leaves the arrow from *A* to *D* intact. Also, our judgment would be that *B* (in all likelihood) did *not* shoot; nothing about *A*'s decision should affect variables in the model that are not effects of *A*'s shot. This bears repeating. If we *see A* shoot, then we conclude that *B* shot too. But if *A decides* to shoot, or if we *make A* shoot, then the opposite is true. This is the difference between *seeing* and *doing*. Only a computer capable of grasping this difference can pass the mini-Turing test.

Note also that merely collecting Big Data would not have helped us ascend the ladder and answer the above questions. Assume that you are a reporter collecting records of execution scenes day after day. Your data will consist of two kinds of events: either all five variables are true, or all of them are false. There is no way that this kind of data, in the absence of an understanding of who listens to whom, will enable you (or any machine learning algorithm) to predict the results of persuading marksman *A* not to shoot.

Finally, to illustrate the third rung of the Ladder of Causation, let's pose a counterfactual question. Suppose the prisoner is lying dead on the ground. From this we can conclude (using level one) that *A* shot, *B* shot, the captain gave the signal, and the court gave the order. But what if *A* had decided not to shoot? Would the prisoner be alive? This question requires us to compare the real world with a fictitious and contradictory world where *A* didn't shoot. In the fictitious world, the arrow leading into *A* is erased to liberate *A* from listening to *C*. Instead *A* is set to false, leaving its past history the same as it was in the real world. So the fictitious world looks like Figure 1.6.

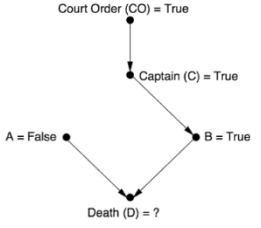


FIGURE 1.6. Counterfactual reasoning. We observe that the prisoner is dead and ask what would have happened if Soldier *A* had decided not to fire.

To pass the mini-Turing test, our computer must conclude that the prisoner would be dead in the fictitious world as well, because *B*'s shot would have killed him. So *A*'s courageous change of heart would not have saved his life. Undoubtedly this is one reason firing squads exist: they guarantee that the court's order will be carried out and also lift some of the burden of responsibility from the individual shooters, who can say with a (somewhat) clean conscience that their actions did not cause the prisoner's death as "he would have died anyway."

It may seem as if we are going to a lot of trouble to answer toy questions whose answer was obvious anyway. I completely agree! Causal reasoning is easy for you because you are human, and you were once a three-year-old, and you had a marvelous three-year-old brain that understood causation better than any animal or computer. The whole point of the "mini-Turing problem" is to make causal reasoning feasible for computers too. In the process, we might learn something about how humans do it. As all three examples show, we have to teach the computer how to selectively break the rules of logic. Computers are not good at breaking rules, a skill at which children excel. (Cavemen too! The Lion Man could not have been created without a breach of the rules about what head goes with what body.)

However, let's not get too complacent about human superiority. Humans may have a much harder time reaching correct causal conclusions in a great many situations. For example, there could be many more variables, and they might not be simple binary (true/false) variables. Instead of predicting whether a prisoner is alive or dead, we might want to predict how much the unemployment rate will go up if we raise the minimum wage. This kind of quantitative causal reasoning is generally beyond the power of our intuition. Also, in the firing squad example we ruled out uncertainties: maybe the captain gave his order a split second after rifleman *A* decided to shoot, maybe rifleman *B*'s gun jammed, and so forth. To handle uncertainty we need information about the likelihood that the such abnormalities will occur.

Let me give you an example in which probabilities make all the

cause the prisoner's death?" or "What are the direct effects of vaccinations?" Had we constructed the diagram by asking about mere associations, it would not have given us these capabilities. For example, in Figure 1.7, if we reversed the arrow Vaccination → Smallpox, we would get the same associations in the data but would erroneously conclude that smallpox affects vaccination.

Decades' worth of experience with these kinds of questions has convinced me that, in both a cognitive and a philosophical sense, the idea of causes and effects is much more fundamental than the idea of probability. We begin learning causes and effects before we understand language and before we know any mathematics. (Research has shown that three-year-olds already understand the entire Ladder of Causation.) Likewise, the knowledge conveyed in a causal diagram is typically much more robust than that encoded in a probability distribution. For example, suppose that times have changed and a much safer and more effective vaccine is introduced. Suppose, further, that due to improved hygiene and socioeconomic conditions, the danger of contracting smallpox has diminished. These changes will drastically affect all the probabilities involved; yet, remarkably, the structure of the diagram will remain invariant. This is the key secret of causal modeling. Moreover, once we go through the analysis and find how to estimate the benefit of vaccination from data, we do not have to repeat the entire analysis from scratch. As discussed in the Introduction, the same estimand (i.e., recipe for answering the query) will remain valid and, as long as the diagram does not change, can be applied to the new data and produce a new estimate for our query. It is because of this robustness, I conjecture, that human intuition is organized around causal, not statistical, relations.

ON PROBABILITIES AND CAUSATION

The recognition that causation is not reducible to probabilities has been very hard-won, both for me personally and for philosophers and scientists in general. Understanding the meaning of "cause" has been the focus of a long tradition of philosophers, from David Hume and John Stuart Mill in the 1700s and 1800s, to Hans