Chapter 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 Adam from eating 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 sides of the story. (In Israeli schools, Genesis is read 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, military conquests, 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 it was reignited suddenly in the 1990s, when I was writing my book *Causality*, and came to confront the Ladder of Causation.

As I re-read 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. The interesting thing, though, is that 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 color. Where did they get this idea?

For me, these nuances carried three profound messages. First, that very early in our evolution, humans came to realize that the world is not made up only of dry facts (what we might call data today), but that these facts are glued together by an intricate web of cause-effect relationships. Second, that causal explanations, not dry facts, make up the bulk of our knowledge, and that satisfying our craving for explanation should be the cornerstone of machine intelligence. Finally, that our transition from processors of data to makers of explanations was not gradual—it required an external push from an uncommon fruit. This matched perfectly what I 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 of course won't find the Tree of Knowledge, but we still see a major unexplained transition. We understand now that humans evolved from ape-like ancestors over a period of 5-6 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 Forward. Humans acquired the ability to modify their environment and their own abilities at a dramatically faster rate.

For example, over millions of years, eagles and owls have evolved truly amazing eyesight—yet they never evolved eyeglasses, microscopes, telescopes, or night-vision goggles. Humans have produced these miracles in a matter of centuries. I call this phenomenon the "super-evolutionary speedup." Some readers might object to my comparing apples and oranges, evolution to engineering, but that is exactly my point. Evolution has endowed us with the ability to engineer our lives, a gift she has not bestowed upon eagles and owls, and the question is again, Why? What computational facility did humans suddenly acquire that eagles lacked?

Many theories have been proposed, but there is one I like because it is especially pertinent to the idea of causation. In his book *Sapiens*, historian Yuval Harari posits that our ancestors' capacity to *imagine* non-existent things was the key to everything, for it allowed them to communicate better. Before this change, they could only trust people from their immediate family or tribe. Afterward their trust extended to larger communities, bound by common beliefs and common expectations (for example, beliefs in invisible yet imaginable deities, in the afterlife, and in the divinity of the leader). Whether you agree with Harari's theory or not, the connection between imagining and causal relations is almost self-evident. It is useless to know the causes of things unless you can imagine their consequences. Conversely, you cannot claim that Eve caused you to eat from the tree unless you can imagine a world in which, counter to facts, she did not hand you the apple.

Back to our *H. sapiens* ancestors: their newly acquired causal imagination enabled them to do many things more efficiently, through a tricky process we call "planning." Imagine a tribe preparing for a mammoth hunt. What would it take for them to succeed? My mammoth-hunting skills are rusty, I must admit, but as a student of thinking machines I have learned one thing. The only way a thinking entity (computer, caveman, or professor) can accomplish a task of such magnitude is to plan things in advance. To decide how many hunters to recruit; to gauge, given

the wind conditions, what direction to approach the mammoth; and more. In short, to imagine and compare the consequences of several hunting strategies. To do this, it must possess, consult, and manipulate a mental model of its reality.

Here is how we might draw such a mental model:

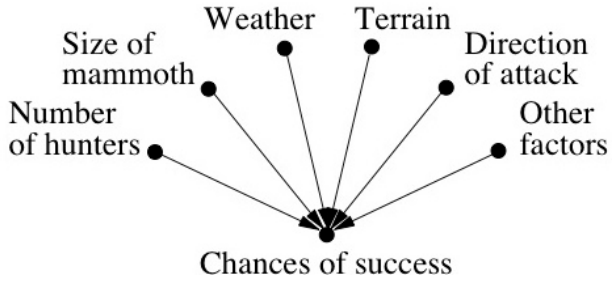


Figure 1. Perceived causes of a successful mammoth hunt.

Each dot in the diagram represents a cause of success. Note that there are multiple causes, and that none of them are deterministic. That is, we cannot be *sure* that more hunters will enable us to succeed, or that rain will prevent us from succeeding; but these factors do change our *probability* of success.

The mental model is the arena where imagination takes place. It enables us to experiment with different scenarios, by making local alterations to the model. Somewhere in our hunters' mental model was a subroutine that evaluated the effect of the number of hunters. When they considered adding more, they didn't have to evaluate every other factor from scratch. They could make a local change to the model, replacing "Hunters = 8" by "Hunters = 9" and re-evaluating the probability of success. This modularity is a key feature of causal models. .

I don't mean to imply, of course, that early humans actually drew a pictorial model like this one. Of course not! But when we seek to emulate human thought on a computer, or indeed when we try to solve unfamiliar scientific problems, drawing an explicit dots-and-arrows picture

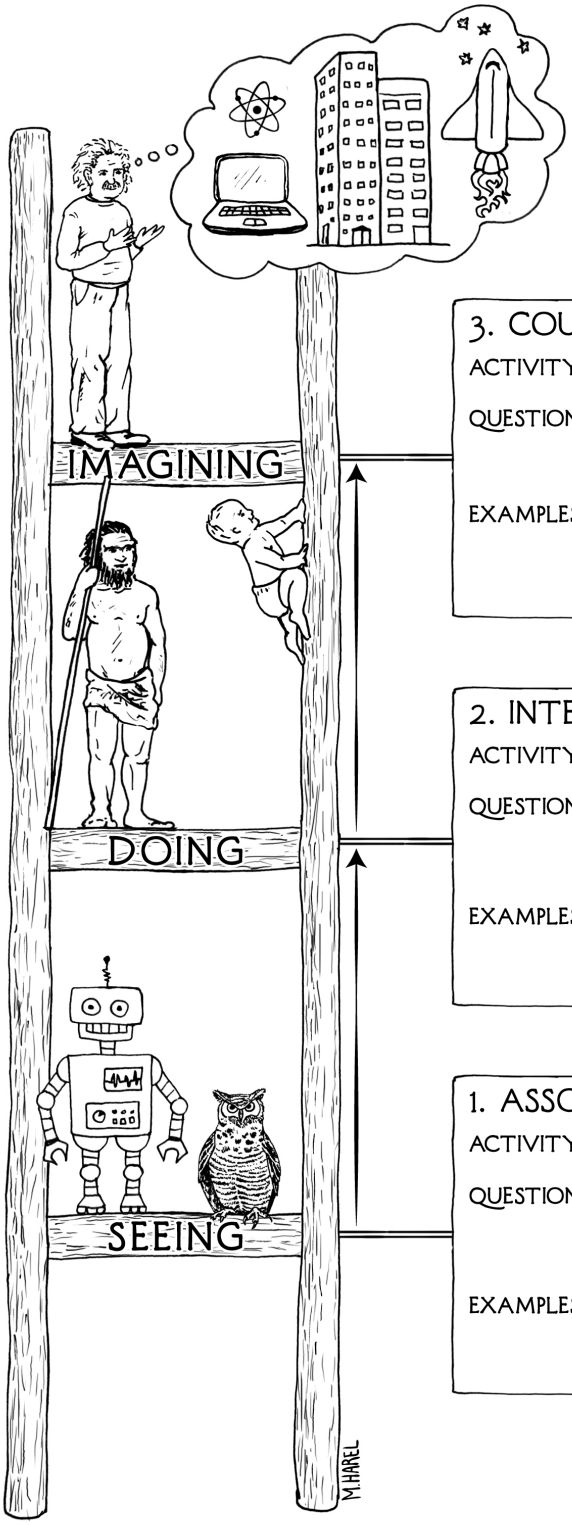
is extremely useful. You will see many in this book, and I will call them causal diagrams. They are the computational core of the "causal inference engine" described in Chapter 1.

The Three Levels of Causation

So far I may have given the impression that the ability to organize our knowledge of the world into causes and effects was monolithic and acquired all at once. But in fact, my research on machine learning has taught me that there are at least three distinct levels that need to be conquered by a causal learner: seeing, doing, and imagining.

The first cognitive ability, seeing or observation, is the detection of regularities in our environment, and it is shared by many animals as well as early humans before the Cognitive Revolution. The second ability, doing, stands for 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. Usage of tools, provided they are designed for a purpose and not just picked up by accident or copied from one's 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 their tool works and what to do when it doesn't. For that, you need to be at 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 planet Earth.

I cannot prove this, but what I can prove mathematically is that the three levels are fundamentally different, each unleashing capabilities that the ones below it do not. The framework I will use to show this goes back to Alan Turing, the pioneer of research in artificial intelligence, who proposed to classify a cognitive system *in terms of the queries it can answer*.



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 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. On the top rung, counterfactual learners can imagine worlds that do not exist and infer reasons for observed phenomena.

This is an exceptionally fruitful approach when we are talking about causality, because it bypasses long and unproductive discussions of "What is causality exactly?" 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? While Turing was looking for a binary classification—human or non-human—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 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 it observes how a rat moves and figures out where it 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 we see* [X]?" For instance, imagine a marketing director at a department store, who asks, "How likely is it that a customer who bought toothpaste will 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|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. A typical measure of association, which we will mention often in this book, is called "correlation" or "regression," which involves fitting a line to a collection of data points and taking the slope of that line. Some associations might have obvious causal interpretations. Other associations 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 their wisdom with an owl. We hear almost every day, it seems, about rapid advances in machine learning systems—self-driving cars, speech-understanding systems, and especially in recent years, deep-learning algorithms (or deep neural networks). How could it be that they are still 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 so difficult. It has not addressed the truly difficult questions that continue to prevent us from achieving human-like AI. The result is that 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 that machines are still disappointingly far from

human-like cognition. My colleague in computer science at UCLA, Adnan Darwiche, has just (as of July 2017) written a position paper called "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 human-like intelligence, able to converse with humans and guide us. What we have gotten from deep learning instead is machines with abilities—truly impressive abilities—but no intelligence. The difference is profound, and lies in the absence of a model of reality.

Just as they did 30 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 still, what drives the fitting process is raw data. They continue to improve in accuracy as more data are fitted, but they do not benefit from the "super-evolutionary speedup" that we encountered above. They end up with a brittle, special-purpose system that is inscrutable even to its programmers. The architects of a program like AlphaGo (which recently defeated the best human go players) do not really know *why* it works, only that it does. The lack of flexibility, adaptability, and transparency is not in the least bit surprising; it is inevitable in any system that works at the first level of the Ladder of Causation.

We step up to the next level of causal queries when we begin to change the world. A typical question for this level is, "What will happen to our floss sales if we double the price of toothpaste?" This already calls for a new kind of knowledge, absent from the data, which we find at rung two of the Ladder of Causation, Intervention.

Intervention ranks higher than Association because it involves not just seeing what is, but changing what is. Seeing smoke tells us a totally different story about the likelihood of fire than

making smoke. Questions about interventions cannot be answered by using passively collected data, no matter how big the data or how deep your neural network. It has been quite traumatic for many scientists to learn that none of the methods they learn in statistics are sufficient even to articulate, let alone answer, a simple question like, "What happens if we double the price?" I know this because I have had many occasions to help them climb to the next rung of the ladder.

Why can't we answer our floss question just by observation? Why not just go into our vast database of previous purchases and see what happened previously when toothpaste cost twice as much? The reason is that on the previous occasions, the price may have been higher for different reasons. For example, the product may have been in short supply, and every other store also had to raise its price. But now you are considering a deliberate intervention that will set a new price regardless of market conditions. The result might be quite different from what it was when the customer couldn't find a better deal anywhere else. If you had data on the market conditions that existed on the previous occasions, perhaps you could figure it out… but what data do you need? And then, how would you figure it out? Those are exactly the questions the science of causal inference allows us to answer.

One very direct way to predict the result of an intervention is to experiment with it under carefully controlled conditions. Big Data companies like Facebook know this, and they constantly perform experiments to see what happens if items on the screen are arranged differently, or if the customer is given a different prompt (or even a different price).

What is more interesting, and less widely known—even in Silicon Valley—is that successful predictions of the effects of interventions can sometimes be made even without an experiment. For example, the sales manager could develop a model of consumer behavior that includes market conditions. Even if she doesn't have data on every factor, she might have data on enough key surrogates to make the prediction. A sufficiently strong and accurate causal

model can allow us to use rung-one (observational) data to answer rung-two (interventional) queries. Without the causal model, we could not go from rung one to rung two. This is why deep-learning systems (as long as they use only rung-one data and do not have a causal model) will never be able to answer questions about interventions, which by definition break the rules of the environment the machine was trained in.

As these examples illustrate, the defining query of the second rung of the Ladder of Causation is, "*What if we do*?" What will happen if we *change* the environment? We can write this kind of query as P(Floss|*do*(Toothpaste)), the probability that we will sell floss at a certain price, given that we *set* the price of toothpaste at another price.

Another popular question at the second level of causation is "How?", which is a cousin of "What if we do?" For instance, the manager may tell us that we have too much toothpaste in our warehouse. "How can we sell it?" he asks. That is, at what price should we set it? Again, the question refers to an intervention, which we want to perform mentally before we decide whether and how to do it in real life. That requires a causal model.

Interventions occur all the time in our daily lives, although we don't usually use such a fancy term for them. For example, when we take aspirin to cure a headache, we are intervening on one variable (the quantity of aspirin in our body) in order to affect another one (our headache status). If we are correct in our causal belief about aspirin, the "outcome" variable will 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 often wish to ask: My headache is gone now, but *why*? Was it the aspirin I took? The food that 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, in which some observed facts are bluntly negated. Yet the human mind does make such explanation-seeking inferences, reliably and repeatably. Eve did it when she identified "The serpent deceived me" as the reason for her action. This is the ability that most distinguishes human from animal intelligence, as well as 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 it 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, on hundreds of springs, in dozens of laboratories and on thousands of different occasions. However, once anointed with the term "law," physicists interpret it as a functional relationship that governs this very spring, at this very moment of time, 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 being 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 from 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 3). The Lion Man, roughly 40 thousand years old, is a mammoth tusk that has been sculpted in 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 it was made by anatomically modern humans 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.

As a manifestation of our newfound ability to imagine things that have never existed, the Lion Man is the precursor to 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.

Figure 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 ability to reason about counterfactuals.

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 on and improve upon past actions and, perhaps even more significant, our willingness to take responsibility for past and current actions.

As shown in Figure 2, the characteristic queries for the third rung of the Ladder of Causation are "What if I had done…?" and "Why?" Both of these involve comparing the observed world to a counterfactual world. Such questions cannot be answered by experiments alone. 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 are extraordinarily confident) a "law of nature." In short, we need understanding.This is, of course, one of the holy grails of any branch of science—the development of a theory that will enable us to predict what will happen in situations we have not even envisioned yet. But it goes even further: having such laws permits us to violate them selectively, so as to create worlds that contradict ours. Our next section will feature such violations in action.

The Mini-Turing Test

In 1950, Alan Turing asked what it would mean for a computer to think like a human. He suggested a practical test, which he called "the imitation game," but every AI researcher since then has called it the "Turing test." For all practical purposes, a computer could be called a thinking machine if an ordinary human, communicating with the computer by typewriter, would not be able to tell whether he was talking with a human or a computer. Turing was very confident that this was within the realm of feasibility. "I believe that in about fifty years' time it will be possible to program computers," he wrote, "… to make them play the imitation game so well that an average interrogator will not have more than a 70 percent chance of making the right identification after five minutes of questioning."

Turing's prediction was slightly off. Every year the Loebner Prize competition identifies the most human-like "chatbot" in the world, with a gold medal and $100,000 offered to any program that succeeds in fooling all four judges into thinking that it is human. As of 2015, in 25 years of competition, not a single program has fooled all the judges or even half of them.

Turing didn't just suggest the "imitation game," he also proposed a strategy to pass it. "Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's?" he asked. If you could do that, then you could just teach it the same way you would teach a child and presto, twenty years later (or less, given a computer's greater speed), you would have an artificial intelligence. "Presumably the child brain is something like a notebook as one buys it from the stationer's," he wrote. "Rather little mechanism, and lots of blank sheets." He was wrong about that: the child's brain is rich in mechanisms and pre-stored templates.

Nonetheless, I think that Turing's instinct had more than a kernel of truth. We probably will not succeed in creating human-like intelligence until we can create child-like intelligence, and a key component in this intelligence is the mastery of causation.

How can machines acquire causal knowledge? That is still a major challenge which undoubtedly will involve an intricate combination of inputs from active experimentation, passive observation, and (not least) input from the programmer—much the same inputs that a child receives, with evolution, parents, and peers substituted for the programmer.

However, we can answer a slightly less ambitious question: How can machines (and people) *represent* causal knowledge, in a way that would enable them to access the necessary information swiftly, answer questions correctly, and do it with ease, as a three-year-old child can? In fact, this is the main question we address in this book.

I call this the mini-Turing test. The idea is to take a simple story, encode it on a machine in some way, and then test to see if the machine can correctly answer causal questions that a human can answer. It is "mini" for two reasons. First, because it is confined to causal reasoning, excluding other aspects of human intelligence such as vision and natural language. Second, we allow the contestant to encode the story in any convenient representation, unburdening the machine from the task of acquiring the story from its own personal experience. Passing this mini-test has been my life's work—consciously for the last twenty-five years, and subconsciously even before that.

Obviously, as we prepare to take the mini-Turing test, the question of representation needs to precede the question of acquisition. Without a representation, we wouldn't know how to store information for future usage. Even if we could let our robot manipulate its environment at will, whatever information we learned this way is destined to be forgotten, unless our robot is endowed with a template to encode the results of those manipulations. One major contribution of AI to the study of cognition has been the paradigm: "Representation first, acquisition second." Often it turned out that the quest for a good representation led to insights on how the knowledge ought to be acquired, be it from data or a programmer.

When I describe the mini-Turing test to people, one common reaction is to 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 it, 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?

This cheating possibility, known as the "Chinese Room Argument," was introduced in 1980 by the philosopher John Searle 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 10 causal variables, each of which takes only two values (0 or 1). There are roughly 30 million possible queries that we could ask, 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 to have more entries than the number of atoms in the universe. So it is clear that a dumb list of questions and answers will never be able to 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 it 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 something like this. But more important for our purposes, these models pass the mini-Turing test; no other model is known to do so. Let's look at some examples.

Suppose that a prisoner is about to be executed by a firing squad. A certain chain of events has to occur for this to happen. First, the court has to order 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.

Here is the diagram representing the story I just told:

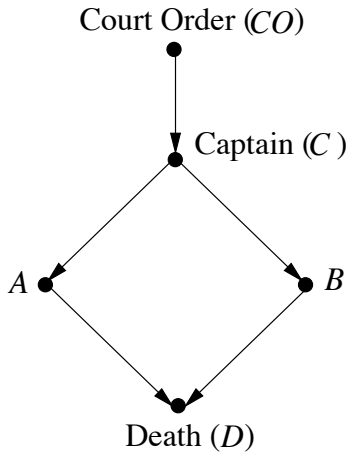


Figure 4. Causal diagram for the firing squad example. A and B represent (the actions of) soldiers A and B.

In Figure 4, 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 ordinary 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 of 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 as compared to *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.

In our example, the resulting causal diagram is shown in Figure 5. Under this intervention, the result is inevitably the prisoner's death. That is the causal meaning of the arrow leading from A to D.

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 A → 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

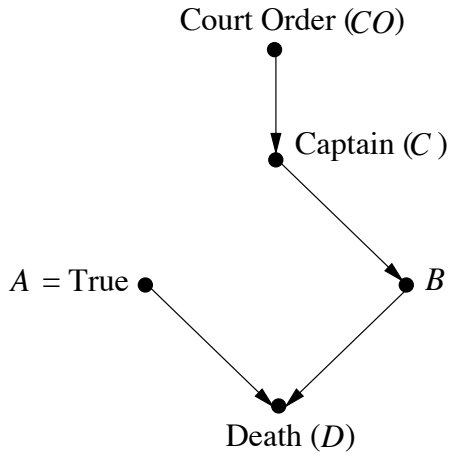


Figure 5. Reasoning about interventions. Soldier A decides to fire; arrow from C to A is deleted and A is assigned the value True.

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[2]. This is the difference between *seeing* and *doing*. Only a computer capable of grasping this difference can pass the mini-Turing test.

[2] Another way to say this is that when evaluating an intervention in a causal model, we make the minimum changes possible to enforce its immediate effect. So we "break" the model where it comes to A, but not B.

Note also that merely collecting big data would not have helped us go up 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 answer 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. If, contrary to fact, 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 and A is set to False, but the past history of A stays the same as it was in the real world. So the fictitious world looks like Figure 6.

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 of the reasons firing squads exist. They guarantee that the court's order will be carried out, and they also lift some of the burden of responsibility off the individual shooters, who can say with a (somewhat) clean conscience that their action did not cause the prisoner's death: "He would have died anyway."

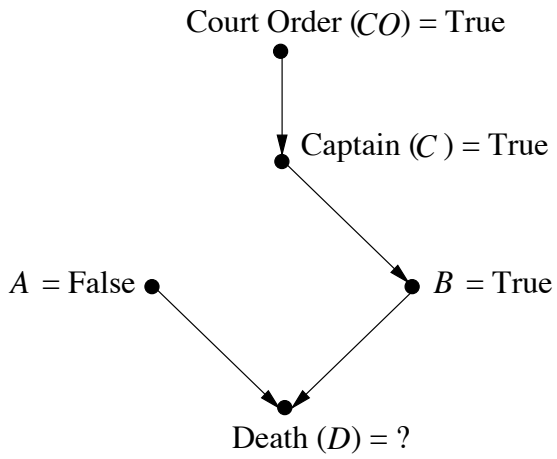


Figure 6. Counterfactual reasoning. We observe that the prisoner is dead, and ask what would have happened if Soldier A had decided not to fire.

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 we have seen in all three examples, 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 breaking the rules about what head goes with what body.)

However, let's not get too complacent about human superiority. There are a great many situations where humans may have a much harder time reaching correct causal conclusions. 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 would go up in the event of a raise in 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, etc. To handle uncertainty we need information on how likely the alternatives are to occur.

Let me give you an example in which probabilities make all the difference. It echoes the public debate that erupted in Europe when smallpox vaccination was first introduced. Unexpectedly, data showed that more people died from smallpox inoculations than from smallpox itself. Naturally, some people used this information to argue that inoculation should be banned when, in fact, it was saving lives by eradicating smallpox. Let's look at some fictitious data to illustrate the effect and settle the dispute.

Suppose that out of 1 million children, 99 percent are vaccinated and 1 percent are not. If a child is vaccinated, he or she has 1 chance in 100 of developing a reaction, and the reaction has 1 chance in 100 of being fatal. On the other hand, he or she has no chance of developing smallpox. Meanwhile, if a child is not vaccinated, he or she obviously has zero chance of developing a reaction to the vaccine, but he or she has 1 chance in 50 of developing smallpox. Finally, let's assume that smallpox is fatal in one out of 5 cases.

I think you would agree that vaccination looks like a good idea. The odds of having a reaction are less than the odds of getting smallpox, and the reaction is much less dangerous than the disease. But now let's look at the data. Out of 1 million children, 990 thousand get vaccinated; 9,900 get the reaction; and 99 die from the reaction. Meanwhile, 10 thousand don't get vaccinated, 200 get smallpox, and 40 die from the disease. In summary, more children die from vaccination (99) than from the disease (40).

I can empathize with the parents who might march to the health department with signs saying, "Vaccines killed our children!" And the data seem to be on their side; the vaccinations

indeed are causing more death than smallpox itself. But is logic on their side? Should we ban vaccination, or take into account the deaths prevented? We can make this clear using the causal diagram shown in Figure 7.

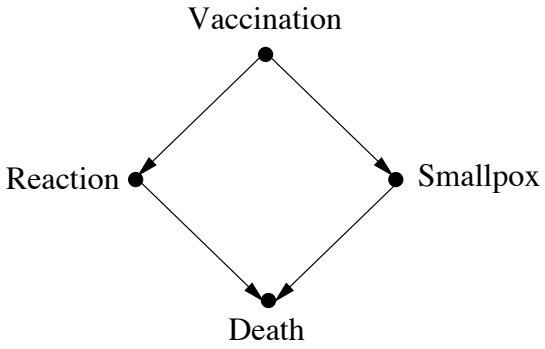


Figure 7. Causal diagram for vaccination example. Is vaccination beneficial or harmful?

When we began, the vaccination rate was 99 percent. We now ask the counterfactual question: What if we had set the vaccination rate to 0? Using the probabilities I gave you above, we can conclude that out of 1 million children, 20 thousand would have gotten smallpox and 4,000 would have died. Comparing the counterfactual world with the real world, we see that the cost of not vaccinating was the death of 3,861 children (the difference between 4,000 and 139). We should thank the language of counterfactuals for helping us to avoid such costs.[3]

The main lesson for a student of causality is that there is more to a causal model than merely writing arrows. Behind the arrows, there are probabilities. When we draw an arrow from X to Y, then implicitly we are saying that there is some probability rule or function specifying how Y would change if X were to change. We might know what the rule is; more likely, we will have to estimate it from data. One of the most intriguing features of the Causal Revolution,

[3] I should also mention here that counterfactuals allow us to talk about causality in individual cases: what would have happened to Mr. Smith, who was not vaccinated and died of smallpox, if he had not been vaccinated? Such questions, the backbone of personalized medicine, cannot be answered from rung-two information.

though, is the fact that we can in many cases leave those mathematical details completely unspecified. Very often the *structure of the diagram itself* enables us to estimate all sorts of causal and counterfactual relationships: simple or complicated, deterministic or probabilistic, linear or non-linear.

From the computing perspective, another remarkable thing about our scheme for passing the mini-Turing test is the fact that we used the same routine in all three examples: Translate the story into a diagram, listen to the query, perform a surgery that corresponds to the given query (interventional or counterfactual; if the query is associational than no surgery is needed), and then use the modified causal model to compute the answer. We did not have to train the machine on a multitude of new queries each time we changed the story. The approach is flexible enough to work whenever we can draw a causal diagram, whether it has to do with mammoths, firing squads, or vaccinations. This is exactly what we want for a causal inference engine: it is the kind of flexibility we enjoy as humans.

Of course, there is nothing inherently magic about a diagram. The success of the diagram is attributable to the fact that it carries causal information; that is, when we constructed the diagram we asked ourselves, "Who could be a direct cause of 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 7, if we reversed the arrow Vaccination → Smallpox we would get the same associations in the data, but we would erroneously conclude that smallpox affects vaccination.

Decades of experience with these kinds of questions have given me a firm conviction that, in both a cognitive sense 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 the knowledge encoded in a probability distribution. For example, suppose that times have changed and a new vaccine is introduced which is much safer and more effective. Suppose, further, that due to improved hygiene and socioeconomic conditions, the danger of contracting smallpox has diminished. These changes will drastically affect all of the probabilities involved, yet, remarkably, the structure of the diagram would 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, it 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. The drive to understand what a "cause" means has been the focus of a long tradition of philosophers, from Hume and Mill in the 1700s and 1800s to Hans Reichenbach and Patrick Suppes in the mid-1900s, to Nancy Cartwright, Wolfgang Spohn and Christopher Hitchcock today. In particular, beginning with Reichenbach and Suppes, philosophers have tried to define causation in terms of probability, using the notion of "probability raising": X causes Y if X *raises the probability* of Y.

This concept is solidly ensconced in intuition. We say, for example, "reckless driving causes accidents" or "you will fail this course because of your laziness," knowing quite well that

the antecedents merely tend to make the consequences more likely, not absolutely certain. One would expect, therefore, that probability raising should become the bridge between rung one and rung two of the Ladder of Causation. Alas, this intuition has led to decades of failed attempts.

What prevented the attempts from succeeding was not the idea itself but the way it was articulated formally. Almost without exception, philosophers expressed the sentence, "X raises the probability of Y" using conditional probabilities and wrote: $P(Y|X) > P(Y)$. This interpretation is wrong, as you surely noticed, because "raises" is a causal concept, connoting a causal influence that X has over Y. The expression $P(Y|X) > P(Y)$, on the other hand, speaks only about observations, and means, "If we see X, then the probability of Y increases." But this increase may come about for other reasons, including Y being a cause of X or some other variable (Z) being the cause of both of them. That's the catch! It puts the philosophers back on square one, trying to eliminate those "other reasons."

Probabilities, as given by expressions like $P(Y|X)$, lie on the first rung of the Ladder of Causation and they cannot ever (by themselves) answer queries on the second or third rung. Any attempt to "define" causation in terms of simpler, first-rung concepts must fail. That is why I have not attempted to define causation anywhere in this book; definitions demand reduction and reduction demands going to a lower rung. Instead, I have pursued the ultimately more constructive program of explaining how to answer causal queries and what information is needed to answer them. If this seems odd, consider that mathematicians take exactly the same approach to Euclidean geometry. Nowhere in a geometry book will you find a definition of the terms "point" and "line." Yet we can answer any and all queries about them on the basis of Euclid's axioms (or even better, the various modern versions of Euclid's axioms).[4]

[4] To be more precise: in geometry, undefined terms like "point" and "line" are primitives. The primitive in causal inference is the relation of "listening to," indicated by an arrow.

But let's look at this criterion of probability raising more carefully and see where it runs aground. The issue of a common cause or *confounder* of X and Y, mentioned above, was one of the most vexing ones for philosophers. If we take the probability-raising criterion at face value, we would have to conclude that high ice cream sales cause crime, because the probability of crime is higher in months when more ice cream is sold. In this particular case, we can explain the phenomenon because both ice cream sales and crime are higher in summer, when the weather is warmer. Nevertheless, the question remains: what general philosophical criterion could tell us that weather is the cause, not ice cream sales?

Philosophers tried hard to repair the definition by conditioning on what they called "background factors" (another word for confounders), yielding the criterion $P(Y|X, K = k) > P(Y|K = k)$, where K stands for some background variables. In fact, this criterion works for our ice cream example, if we treat temperature as a background variable. For example, if we look only at days when the temperature is 90 degrees (K = 90), we will find no residual association between ice cream sales and crime. It's only when we compare 90-degree days to 30-degree days that we get the illusion of a probability raising.

Still, no philosopher has been able to give a convincingly general answer to the question: Which variables need to be included in the background set K and conditioned on? The reason is obvious; confounding too is a causal concept, hence defies probabilistic formulation. In 1983, Nancy Cartwright broke this deadlock and enriched the description of the background context with a causal component. She proposed that we should condition on any factor that is "causally relevant" to the effect. By borrowing a concept from rung two of the Ladder of Causation she essentially gave up on the idea of defining causes from probability alone. This was progress, but it opens the door to the criticism that we are defining a cause in terms of itself.

Philosophical disputes over the appropriate content of K continued for more than two decades and reached an impasse. In fact, we will see a correct criterion in Chapter 4 and I will not spoil the surprise here. It will suffice for the moment to say that this criterion is practically impossible to enunciate without causal diagrams.

In summary, confounding has always been the rock on which probabilistic causality has foundered. Every time the adherents of probabilistic causation try to patch up the ship with a new hull, the boat runs into the same rock and springs another leak. Once you misrepresent "probability raising" in the language of conditional probabilities, no amount of probabilistic patching will get you to the next rung of the ladder. As strange as it may sound, the notion of probability raising cannot be expressed in terms of probablities.

The proper way to rescue the probability-raising idea would be with the *do*-operator: we could say that X causes Y if $P(Y| do(X)) > P(Y)$. Since intervention is a rung-two concept, this definition can capture the causal notion of probability raising, and it can also be made operational through causal diagrams. In other words, if we have a causal diagram and data on hand and a researcher asks whether $P(Y| do(X)) > P(Y)$, we can answer his question coherently and algorithmically, and thus decide if X is a cause of Y in the probability-raising sense.

I usually pay a great deal of attention to what philosophers have to say about slippery concepts such as causation, induction, and the logic of scientific inference. Philosophers have the advantage of standing apart from the hurly-burly of scientific debate and the practical realities of dealing with data. They have been less contaminated than other scientists by the anti-causal biases of statistics. They can call upon a tradition of thought about causation that goes back at least to Aristotle, and they can talk about causation without blushing or hiding it behind the label of "association."

However, in their effort to mathematize the concept of causation—itself a laudable idea—philosophers were too quick to commit to the only uncertainty-handling language they knew, the language of probability. They have for the most part gotten over this blunder, but unfortunately similar ideas are being pursued in econometrics even now, under names like "Granger causality" and "vector autocorrelation."

Now I have a confession to make: I made the same mistake, too. I did not always put causality first and probability second. Quite the opposite! When I started working in artificial intelligence, in the early 1980s, I thought that the *most important* thing missing from AI's was uncertainty. Moreover, I was insisting that uncertainty be represented by probabilities. Thus, as I will explain in Chapter 3, I developed an approach to reasoning under uncertainty, called Bayesian networks, which mimics how an idealized, decentralized brain might incorporate probabilities into its decisions. Given that we see certain facts, Bayesian networks can swiftly compute how likely it is that certain other facts are true or false. Not surprisingly, Bayesian networks caught on immediately in the AI community, and even today they are considered one of the leading paradigms in artificial intelligence for reasoning under uncertainty.

Though I am delighted with the ongoing success of Bayesian networks, they failed to bridge the gap between artificial and human intelligence. I'm sure you can figure out the missing ingredient: causality. True, causal ghosts were all over the place. The arrows invariably pointed from causes to effects, and practitioners often noted that diagnostic systems became unmanageable when the direction of the arrows was reversed. But for the most part we thought that this was a cultural habit, or an artifact of old thought patterns, not a central aspect of intelligent behavior.

At the time, I was so intoxicated with the power of probabilities that I considered causality to be a subservient concept, merely a convenience or a mental shorthand for expressing

probabilistic dependencies, and for distinguishing relevant variables from irrelevant ones. In my 1988 book *Probabilistic Reasoning in Intelligent Systems*, I wrote, "Causation is a language with which one can talk efficiently about certain structures of relevance relationships." The words embarrass me today, because "relevance" is so obviously a rung 1 notion. Even by the time the book was published, I knew in my heart that I was wrong. To my fellow computer scientists, my book became the bible of reasoning under uncertainty, but I was already feeling like an apostate.

Bayesian networks inhabit a world where all questions are reducible to probabilities, or (to put it in the terminology of this chapter) degrees of association between variables; they could not ascend to the second or third rungs of the Ladder of Causation. Fortunately, they required only two slight twists to climb to the top. First, in 1991, the graph surgery idea empowered them to handle both observations and interventions. Another twist, in 1994, brought them to the third level and made them capable of handling counterfactuals. But these developments deserve a fuller discussion in a later chapter. The main point is this: While probabilities encode our beliefs about a static world, causality tells us whether and how probabilities change *when the world changes*, be it by intervention or by act of imagination.