TECHNICAL REPORT
R-536
October 2025

Preface to the Second Edition

Eight years have passed since I wrote the preface to the first English edition of *The Book of Why*, and in that time major changes have taken place in the fields of statistics, social and health science, artificial intelligence (AI) and machine learning (ML). Most importantly, causal inference, a revolution that I called a "new science" in the subtitle eight years ago, has taken enormous strides towards general acceptance in those fields. If I were to write the book today, I might not even need the word "new" in the subtitle.

This book is simply about the science of cause and effect. In this brief preface, I will update readers on some trends in causal inference since the first edition that I consider particularly impressive.

Causal Inference: A Revolution Embraced

Causal inference has increasingly been recognized as an independent and essential component in every aspect of intelligent decision making. One visible evidence of this recognition was the 2021 Nobel Prize in Economics, awarded to Guido Imbens and Joshua Angrist for "their methodological contributions to the analysis of causal relationships." Other major awards to causal inference research were given in AI and statistics. In the same year, I received the BBVA Frontiers of Knowledge Award for "laying the foundations of modern AI." A year later, the 2022 Rousseeuw Prize for Statistics was awarded to James Robins, Thomas Richardson, Andrea Rotnitzky, Miguel Hernán, and Eric Tchetgen Tchetgen, for their "pioneering work on causal inference with applications in medicine and public health."

Awards aside, let's examine the science. It is not a secret that I have been critical of the "potential outcomes" framework that Imbens and Angrist take toward causation (elaborated in Chapter 9). Briefly, this framework instructs researchers to accept certain assumptions stated in an opaque language, which prevents them from judging the plausibility of those assumptions. Causal diagrams, by contrast, the language I advocate in this book, empower researchers to state their assumptions explicitly and meaningfully, so that their plausibility can be assessed and scrutinized by the scientific community. Unfortunately, diagrams became taboo in the potential outcomes culture. Imbens and Donald Rubin, for example, in their 2015 textbook, stated, "we have not found this [causal diagrams] approach to aid drawing of causal inferences," and did not include even a single diagram. Earlier, Rubin (2009) dismissed causal diagrams as "non-scientific ad-hockery."

In view of this past history, I was delighted to see a review article by Imbens, published in 2024, which includes diagrams on nearly every page. I take it as evidence that economists in the potential outcomes camp have recognized the unique power of causal diagrams and are now ready to lift the embargo. I would still like to see more recognition of graphs not only as a pedagogical tool, but as a powerful inference tool that facilitates solutions to problems of data fusion, missing data, selection bias, and mediation analysis, problems to which causal diagrams have spawned complete and transparent solutions (Chapters 9 and 10).

Putting aside the issues of methodology, other signs of the causal revolution are all around us. The number of citations to key articles in causal inference research has skyrocketed from 6,000 in 2018 to more than 12,000 in 2024. Dozens, if not hundreds, of seminars, workshops and symposia have been organized to disseminate progress in causal inference methods and applications.

The evidence is not only in numbers, but in words and deeds. In 2024, the American Medical Association announced an epochal change in the way it treats statements about causality in its flagship publication, *Journal of the American Medical Association* (*JAMA*). For comparison purposes, in 2017, the year before *The Book of Why* was published, the editors wrote that when evaluating a submission, they would "verify whether the study is a randomized clinical trial or a report of a controlled laboratory experiment. If it isn't, and is a report of an observational study, … then all cause-and-effect language must be replaced."

This is very much the "old" attitude toward causation, which reigned from 1900 to 2018. But the tide was turning. By 2024, even the august *JAMA* was ready to change its stance. In the June 4 issue, the editors announced that they would now accept papers that draw conclusions about causes and effects even from observational studies, when justified by appropriate causal models. "So with excitement and trepidation," they wrote, "we will now consider how best to balance methodologic advances … with the principles in our long-standing and often discussed reporting policy that generally limits use of causal language to well-done randomized clinical trials."

For readers coming to this topic for the first time, both the excitement and the trepidation may seem a little bit bizarre. Why so much angst over something so obvious as causes and effects? In Chapters 2-5, I try to explain why scientists developed such an extreme aversion to talking about causality. I attribute the problem to the absence of a formal language for expressing causal relationships.

Nevertheless, your surprise is well justified. The ability to understand causes and effects is one of the hallmarks of scientific thinking, and it is unfortunate that the paragons of science neglected this fundamental mode of thinking for so long. At the same time, causality has also been also one of the Achilles heels of artificial intelligence. And that brings us to our next trend…

Causal Inference and Large Language Models: Faking It or Making It?

For many years, a great tension has existed between two competing paradigms of data science, the "data fitting" paradigm on the one hand and the "data interpreting" paradigm on the other. The data-fitting school, which still dominates statistics and machine learning, is driven by the faith that data alone can guide us to rational decisions, if only we are sufficiently clever at data mining. In contrast, the data-interpreting school, which guides causal modeling and forms the backbone of this book, views data not as a sole object of inquiry but an auxiliary means for interpreting reality. In this case, "reality" stands for the processes that generate the data.

Now we understand that these two paradigms can live in symbiotic harmony. Causal modeling

uses qualitative assumptions about reality and aims to identify the quantitative properties of the data needed to answer causal questions. Machine learning, on the other hand, takes us in the opposite direction, from the data available to estimates of those properties that causal analysis identifies as needed. The two approaches converge in the middle.

A key to this symbiosis was the discovery that causal questions form a three-layer hierarchy, which we call the Ladder of Causation (Chapter 1) and refer to throughout this book. Each layer answers a different type of question: questions about association and prediction on rung one, questions about interventions on rung two, and counterfactual questions on layer three. Each layer can do things that the layer below it cannot do. For example, layer 1, which contains knowledge about associations, cannot answer questions about interventions -- the province of layer 2. To answer questions on a higher layer, we require assumptions and data appropriate to that layer. Population-level data, suited to rung two, cannot answer questions about individual people (such as, what would have happened if we had given drug B to patient A who actually took drug B' and recovered?).

In Chapter 10 of *The Book of Why*, I turned my attention to machine learning. As you will see, I was highly skeptical of the ability of then-current machine-learning programs to achieve what is called "strong AI" or "artificial general intelligence," a machine intelligence capable of functioning autonomously in the real world and communicating with humans in natural human language. In particular, I saw the Ladder of Causation as a major stumbling block, because all of the programs then in existence operated on the level of data fitting. They therefore in principle could not ascend past the first rung of the Ladder of Causation.

As almost every reader of this book probably knows, the advent of Large Language Models (LLM), such as ChatGPT and GPT-4, has completely changed the landscape of machine learning. These programs not only simulate human conversations well enough to fool a human, they pass the causal version of the Turing test discussed in Chapter 1. While previous iterations of "AI" failed ignominiously, LLMs can now answer the benchmarked questions in Chapter 1 articulately and correctly.

Does this disprove the Ladder of Causation? Emphatically not. These programs circumvent the limitations of the Ladder by changing the nature of the training data. Instead of training themselves on observations obtained directly from the environment, they are trained on linguistic texts written by human authors who already have well-developed causal models of the world. The programs can simply cite information from the text without experiencing any of the underlying data. The result is a sequence of linguistic extrapolations which, in some remote and obscure sense, reflect the causal understanding of those authors.

With proper prompting and expert guidance, LLMs can now solve some of the toy examples discussed in *The Book of Why*, such as the Firing Squad and the Oxygen-Match-Fire examples. Their solutions give the impression that they have a causal model of the world, when in fact they are merely quoting word sequences from the training texts.

Some researchers have argued that the capacity to fake a causal model is equivalent to actually having one. But is it so?

The answer is not easily discernible when it comes to machines that can store trillions of symbolic sequences. We humans cannot function without a world model because we cannot store the answers to all possible queries and look them up when needed. Instead, we store a parsimonious model of the world and derive the requested answers from the model. Possibly an organism with a gigantic memory may not need to resort to such tricks. Or possibly, it is our human bias toward causal explanations that makes us confuse ersatz causal reasoning with the real thing.

Let me expand on this point, because it leads to two reasons why human-style causal models remain indispensable even in the era of LLM.

First, the art of "prompting" LLM systems is still black magic. Currently, these systems require extremely delicate prompting in order to get the correct answer to a causal query. They occasionally produce unintended "hallucinations" as a result of improper prompting or the presence of questionable data sources. The principles of causal modeling could guide us toward a systematic understanding of the behavior of these programs, thus turning the art of prompting into a science.

Second, in order to be deemed "trustworthy," an LLM must explain its reasoning process to a human user, who will still require an explanation that uses concepts and principles that are compatible with a human-style model of reality. Alternatively, a human expert might be able to explain the LLM to other humans. The latter approach requires a level of transparency that current LLMs do not possess. The former would require the machine to be fluent in a human-style causal model. Either way, human-style causal inference, using the principles explained in this book, will be essential for the "human in the loop" or for the machine itself.

Worldviews, Mindsets, and Narratives: Hidden Nuggets of Causal Inference

Readers often ask me to pin down the single most important contribution of causal inference to AI and to science in general. I have to pause for a moment, because there are so many precious nuggets to choose from, but I usually say, "The mathematization of counterfactuals." (See Chapters 8-9.) Indeed, considering that counterfactuals are indispensible building blocks of scientific thinking, moral judgment and legal arguments (as explained in those chapters), the ability to capture this mode of thought in mathematical notation and infer consequences is an achievement that cannot be understated. I am mighty proud to be part of this development.

However, as I've gained experience in explaining *The Book of Why* to various audiences, I came to realize that there are other nuggets shining modestly throughout the book, whose importance to AI and scientific thought is perhaps even more universal than counterfactuals. I am referring here to the notion of "worldview," also known as "mindset," "narrative," "belief-set," "framework," "paradigm," and many other names that are at the foundation of many disciplines, from literature to politics to health science.

For now, I'll use the word "narrative." A narrative frames reality and distinguishes data from their interpretations -- seeing from understanding. This is very reminiscent of what causal models

do, since a causal model is a special form of a narrative, defined over variables. Other kinds of narratives may be defined in terms of interactions among agents, or distances between "possible worlds," as in Chapter 8.

Two rational agents with different narratives, exposed to the same data, may easily clash in crucial decision-making situations. For a current example, consider the war in Gaza. The Palestinians view Israel as a white-settler project aimed at displacing them. The Israeli narrative, on the other hand, sees that same project as a homecoming endeavor— besieged and harassed by its neighbors' rockets. Religious wars represent an even more extreme case, where minute differences of mythological metaphors have led nations into centuries of death and destruction.

This book is replete with examples where different models of the world led scientists to different decisions, some beneficial and some misguided. Most of these examples focus on medical decision making. But to conclude this Preface, I want to emphasize the universality of the concept of "narrative," and my realization that the mathematical machinery developed in the book constitutes a "mathematization of narratives," not merely of counterfactuals, including narratives involving multiagents motivated by intricate webs of intentions and desires. The mathematical machinery can therefore be used to manage and understand the dynamics of narratives, their formation, mutation, propagation and implications, in a variety of fields. More importantly, the mathematization of narratives allows us to identify the testable implications of each narrative, to search systematically for evidence that would anchor beliefs to reality, and to explore tipping points where non-conforming evidence overturns a prevailing narrative, giving rise to a paradigm shift.