Preface to the French Translation of *The Book of Why*[*]

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, USA
judea@cs.ucla.edu

Almost 7 years have passed since I wrote the preface to the first English edition of *The Book of Why*, and I feel both humble and honored to see the book translated into French, the language of Pascal, Viète, Descartes and Laplace – the giants that shaped scientific methodology.

The past seven years have seen major developments in artificial intelligence (AI), machine learning (ML) and the science of causal modeling. Remarkably, the period has also spawned a greater understanding of how the latter two can be fused together in the service of Artificial General Intelligence (AGI).

When *The Book of Why* was first published (2018), a great tension existed between two competing paradigms, 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 ML, is driven by the faith that the secret to rational decisions lies in the data itself, if only we are sufficiently clever at data mining. In contrast, the data-interpreting school, which guides causal modeling, views data, not as a sole object of inquiry but as an auxiliary means for interpreting reality, and "reality" stands for the processes that generate the data.

We now understand how these two paradigms can work in symbiotic harmony. Causal modeling takes us from qualitative assumptions about reality to properties of the data needed to answer questions about reality. ML on the other hand, take us from the data available toward the best estimates of those properties that causal analysis identifies as needed.

Key to this symbiosis was the discovery that causal questions form a 3-layer hierarchy, as described by the Ladder of Causation (Chapter 1). Each layer answers a different type of question and each can do things that the layer below it cannot do. This limitation informs us what type of data must be gathered to answer a query of a given type.

[*]Pearl, J. and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. NY: Basic Books, 2018.

In the causal modeling arena, the area that has seen the most transformative development in the past seven years has been counterfactual-based decision making, for example in personalized medicine. Many current health-care methods and procedures are guided by population data, obtained from controlled experiments or observational studies. However, the task of going from these data to the level of individual behavior requires counterfactual logic (Chapter 8), where significant new results have been obtained lately.

To exemplify these results, consider the problem of prioritizing patients who are in "greatest need" for treatment, or for testing, or for other scarce resources. "Need" is a counterfactual notion (i.e., patients who would have gotten worse had they not been treated) and cannot be captured by statistical studies alone, be they observational or experimental. A related notion is that of a "harmed" patient, namely, a patient who would die if treated and recover if not treated. Remarkably, despite the individualized character of these notions and the impossibility of observing the same patient both under treatment and under no-treatment, recent developments demonstrate that the "probability of harm" can be quantified by combining data from both experimental and observational studies. (See `https://ucla.in/39Ey8sU`.) The ramifications of these results are enormous, with applications in medicine, marketing and politics, since the essential criterion in every decision making context is always "situation specific," be the situation a patient, a physician, an instrument, a location or a time.

In machine learning, meanwhile, the most significant development came in the form of Large Language Models (LLM), embodied in programs such as ChatGPT and GPT-4. These programs have changed the landscape of the "data fitting" paradigm which, according to the Ladder of Causation, could not correctly answer queries about interventions and counterfactuals unless supplemented with causal knowledge, external to the data.

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 authors who already have 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, LLM programs can now solve some of the canonical toy examples discussed in *The Book of Why*, such as the Firing Squad or the Oxygen.Match-Fire examples. Their solutions give the impression that they have a causal model of the world, when in fact, they merely quoted word sequences from the training texts.

Some researchers have argued nevertheless that the capacity to fake a causal model is equivalent to actually having one, concluding that the models and algorithms described in The Book of Why are not needed. But is it so?

The answer is not easily discernible when it comes to machines that can store trillions of symbolic sequences. We, humans, can't function without a world model because we can't store the answers to all possible queries and just look them up

when needed. Instead, we store a parsimonious model of the world and derive the answers, rather than look them up. In contrast, an organism with a gigantic memory may not need to resort to such tricks.

I would like to conclude this Preface by citing two reasons why human-like 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 weird data sources. The principles of causal modeling should 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 program must explain its reasoning process to a human user, for whom the adequacy of an "explanation" amounts to compatibility with a coherent causal model of reality. It is imperative therefore for an LLM program to be aware of the structure of the models that users employ in understanding the world around them.

Unfolding the intricacies of this structure is the topic of *The Book of Why*, and I hope this translation into French will assist readers to sail the next decade of AI developments.

Judea Pearl
Los Angeles, August, 2024