Response to the book: “*The Book of Why: the new science of cause and effect*”, by Judea Pearl and Dana Mackenzie

> *“Data do not understand causes and effects; humans do.*
> *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.”*

“*The Book of Why: the new science of cause and effect*”, published in May of this year, is possibly the first book written for a lay audience by Turing Prize winner Judea Pearl, who has greatly influenced the field of Artificial Intelligence and said to have pioneered the “Causal Revolution” in statistics. His previous influential work “*Causality*”, targeted at a more specialized audience, offered a rich discussion of the substantial aspects in probabilistic reasoning, many of which are referenced but not discussed at a technical level in the present book: from causal diagrams and the identification of causal effects, to counterfactuals and probability of causation to identification of sufficient and necessary causes. *The Book of Why* focuses on “causal inference” as a process that is key to the human brain’s natural ability to manage causes and effects in our environment, and, the authors argue, the basis for our unique form of intelligence. It is the quality that leads us to constantly ask questions such as “what would have happened in conditions had been different?” – and, consequently, to practice retrospection, imagine alternative realities, and ultimately exercise true understanding. In the era of the “data revolution” and given the expanding use of Big Data to tackle a diversity of scientific and social problems, the authors aim to bring discussion of causal reasoning to the forefront; and, hopefully, to encourage significant changes in how such relationships are treated and investigated in a range of fields. Emphasis is placed on the sparse existence of useful language, symbols and systematic foundation to empirically explore direction of causality (as opposed to *association*) between real-world events throughout most of the history of science. This discussion will first explore the model of causal reasoning that is the foundation of this book’s premise, as well as its practical significance, expanding on its implications for the progress towards strong AI.

Firstly, it is worth characterizing the authors' vision of strong AI, so we may appreciate the significance of Pearl's work on causality for this field. The authors elaborate on the fact that, despite the recent impressive advances in self-driving cars, computer vision, speech-recognition systems, and especially deep learning techniques as a whole, present-day AI is far from humanlike cognition. On one hand, it has succeeded is in showing that certain questions that we thought were difficult in fact were not; on the other, it has not achieved in conquering other key questions that continue to set humans apart from machines. A fitting example is the ability to ask and answer questions that pertain to "*why*" an event takes place. Alan Turing, the pioneer of research in AI, proposed to classify a cognitive system in terms of the queries it can answer. This approach is particularly useful in this case because it "*focuses 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?*" (p. 27). Humans have the natural ability to predict outcomes given a series of events. Pearl asserts that the connection between humans' ability to apply imagination and understanding causal relations are intricately related. Put another way, it would be meaningless to wonder about the causes of a certain event unless we were able to devise some sort of prediction of its consequences. This mechanism is best reflected in the process of planning: it involves estimating the requirements to accomplish a desired task, gauging possible adverse conditions, and mentally comparing possible outcomes based on a number of known variables. We make mental alterations of possible scenarios and find ways to manipulate these variables in ways that raise our probability of achieving our goal. We will now see exactly how this process of *imagining* alternative scenarios is critical in human intelligence.

The key framework used to describe the stages of causal reasoning in this book is the *Ladder of Causation*. Each of the three rungs of the ladder is defined by the increasingly sophisticated causal statements that can be asked and answered at each stage. The first and most primal of these levels, *Association*, is marked by an ability to note the regularities of a given event, and to detect irregularities when they happen. For example, an algorithm that "observes" millions of games of Go and detects which moves are associated with a higher number of wins operates at an associative level of causal reasoning. This form is also reflected in statistical analyses of conditional probability, which evaluate the likelihood of observing an event given another one has taken place. Interestingly, the authors point out that even deep learning algorithms rely on essentially an associational form of learning: they are able to "fit a function" to a stream of observations on a set of data; but, given significant changes in other variables that affect the data, that algorithm might need to be retrained to make accurate predictions. Despite the excitement surrounding deep learning techniques in recent times over successful applications, many AI researchers object to the lack of transparency and traceability of their exact mechanism. Understanding of deep learning is thus purely empirical at this point in time, and offers no guarantees for future success. Here is where the debate over the "black box" approach to AI is fueled[1]. In short, according to Pearl, data-driven approaches in machine learning and Big Data will continue to reside in this rung, as long as they fail to apply a causality-driven model and integrate pre-existing causal knowledge about the variables involved.

Moving up the Ladder of Causation, the second rung is referred to as *Intervention*. It is characterized by an ability to predict the effect of deliberate alterations in the environment. In scientific inquiry, a valid way to causally explain the result of an intervention is to carry out experiments under carefully controlled conditions. With passively collected data, we know that even where a variable of interest holds the desired value, other intervening factors might be at work. Managing and accounting for *confounding, or "lurking third* variables*"* is as cumbersome as commonplace in all areas of science. Randomized controlled trials are indeed the mode of knowledge gathering that are most free from external influences, as they allow us to manipulate one variable in a causal network and observe the changes downstream. Except in cases where a useful randomized controlled trial is achieved, using the term "cause" in statistical work is essentially taboo. However, in many occasions such as research of public health hazards (i.e., investigating the connection between smoking and lung cancer), trials are impractical or simply unethical. Moreover, this stage does not in itself entail a level of understanding where a comprehensive theory of the relationship can be developed. We may still be unable to answer

[1] In 2017, the Future of Life Institute carried out a conference on artificial intelligence and agreed on a set of twenty-three principles for future research in "beneficial AI." One of the resulting recommendations stated that "*If an AI system causes harm, it should be possible to ascertain why*". This point clearly speaks to the importance of transparency a potentially ethical concern.

some interesting questions, such as exactly *why* an outcome was observed, or how to interpret results when they do not match up with previous observations.

This brings us to the third and highest rung, where *Imagination* is introduced. This stage is marked by the use of *counterfactual scenarios* to pose the question of what would have happened if the presumed "causal" condition had not taken place[2]. In this form of reasoning, we may hypothesize that outcome Y would not have taken place had intervention X not been applied. Neither associative statistics nor controlled experiments are equipped to answer such questions. Interrogating a scenario where Y did not happen essentially contradicts the fact that Y did happen. As Pearl concisely states: "*No experiment in the world can deny treatment to an already treated person*" (p. 33). Thinking in terms of counterfactuals allow us the flexibility to reflect and improve upon past actions, and further, to acknowledge responsibility for them. They are also responsible for our ability to engage in philosophical theory, scientific discovery, and technological innovation. In fact, all explorations in these areas, by necessity, first take shape in someone's imagination. Importantly though, counterfactual statements lie at the core of causality and are thus worth a deeper look.

The Scottish philosopher David Hume first included a mention of counterfactuals in his definition of causality. In "*An Enquiry Concerning Human Understanding*", he expressed that "*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*". The first part of this definition (specifically, that "*all the objects, similar to the first, are followed by objects similar to the second")* describes the regularity aspect to a causal relationship: whenever X occurs, Y will be consistently expected to follow. He then adds an interesting element, the essential counterfactual definition: "*if the first object had not been, the second never had existed*". Despite this novel assertion, other

[2] In his paper "*Probabilities of causation: Three counterfactual interpretations and their identification*", Judea Pearl argues that the term "counterfactual" is a misnomer, as it "connotes a statement that stands contrary to facts or, at the very least, a statement that escapes empirical verification. Counterfactuals are in neither category; they are fundamental to scientific thought and carry as clear an empirical message as any scientific law." Further, he states that "the standard counterfactual definition of causation (i.e., that E would not have occurred if it were not for C), captures the notion of 'necessary cause'". His work elaborates on how necessary and sufficient cause should be jointly invoked in the construction of causal explanations in diverse areas such as policy analysis, AI, and psychology.

philosophers mostly bypassed the idea, as the worlds that *could have been* seemed much too elusive to formally explore. However, philosopher David Lewis claimed in his 1973 book “*Counterfactuals*” that the second part of Hume’s definition was actually the most crucial, and could even be kept as the full definition of causality, without regard for regularity. In fact, Lewis offered a well-known counterfactual analysis of causation, by expressing possible scenarios as real concrete entities on par with the actual world. He construed events as “*classes of possible spatiotemporal regions*”, and established a relation of comparative similarity between worlds, where any two worlds could be ordered with respect to their closeness to the actual world.

Although these are very interesting philosophical postulates, it is important for our purposes to bring them back into the context of AI. How would a better target model of AI in formal terms look like according to Pearl? How do counterfactuals concretely fit into this model? In their closing discussion in Chapter 10, the authors express the algorithm for answering causal and counterfactual queries as: “*the simpler probability P(YX = x1 = y’ | X = x), where the machine observes an event X = x but not the outcome Y, and then asks for the outcome under an alternative event X = x’. If it can compute this quantity, the machine can treat its intended action as an observed event (X = x) and ask, “What if I change my mind and do X = x’ instead?*”. On a technical report published last July, Pearl concisely presents the mathematical framework of Structural Causal Models (SCM), a form of “inference engine” which is at work in formalizing counterfactual reasoning within graphical representation. SCMs are comprised of *graphical models*, which serve as the language to represent known facts; *counterfactual logic*, the channel to articulate what we wish to know; and *structural equations*, which serve to tie the two together in a coherent semantics. This, the author holds, is one of the core achievements of the Causal Revolution.

Specifically, why should we be concerned with causal reasoning in AI? There is general agreement that, given current designs, we reach a brick wall if we expect artificially intelligent machines to be able to “speak our language”. That is, if we presume them to effectively grasp human flows of reasoning and hold similarly rich exchanges with us. For instance, how would a training scenario work for a cleaning robot to learn it must not start the vacuum cleaner while someone is asleep (of course, given it could identify the presence of such a condition in its environment)? How flexible and effective could its decision-making process be without understanding that sleeping requires silence, and vacuums create noise, thus it must adapt its behavior and postpone its task? Further, we can argue that, for humans, applying certain forms of counterfactual reasoning ("If I had have done X I could have achieved Y") leads to a powerful form of *learning,* too. Without a doubt, currently a very different form from that of machines.

At this point, it is interesting to take a wider look at how other researchers in the field evaluate the achievements and shortcomings of current AI models. On one hand, Geoffrey Hinton, one of the precursors of deep learning and famous for his groundbreaking work on *neural backpropagation*, remains unsurprisingly optimistic about the track that AI is on. Indeed, efforts to commercialize AI technologies continue to grow as they harness the vast amounts of data available that deep learning so much thrives on. On the other hand, Gary Marcus, a professor of cognitive psychology at NYU who used to direct Uber's AI Lab, recently published a detailed critical appraisal of deep learning as the leading-edge of current AI trends. In it he echoes the belief that it "*is only part of the larger challenge of building intelligent machines*"; although deep learning successfully captures complex correlations between input and output features, it lacks ways of representing causal relationships, and often faces challenges in acquiring abstract ideas. Without new approaches, he worries AI may run into a wall; essentially, all those pressing problems that pattern recognition is not best suited to address. Such concerns seem to not be directly denied even by the strongest supporters of current models, who might instead focus on their confidence that with enough data and computational power, AI will eventually break all such boundaries.

As a final thought, it is important to note that the value of data mining techniques, classical statistical methods and data itself must not be understated. There are many causal questions that might never have been asked if it weren't for methods that allowed us to summarize and transform the data available to develop a promising hypothesis. The work we have analyzed today proposes a framework that would allow models of AI to go a step further, to determine the nature of causal pathways and manage any confounders that might exist. The struggles to develop true intelligent machines and the shortcomings that have been discussed reveal that what we know about intelligence is minuscule compared to the vastness of what is still left to discover.

References:

Hume, D., & Millican, P. F. (2007). An enquiry concerning human understanding. Oxford: Oxford University Press.

Marcus, G. (2018). Deep learning: A critical appraisal. https://arxiv.org/pdf/1801.00631.pdf

Pearl, J. (1999). Probabilities of Causation: Three Counterfactual Interpretations and their identification. *Synthese, 121*, 93-149.

Pearl, J. (2000). Causality: Models, reasoning and inference. Cambridge: Cambridge University Press

Pearl, J. (2018). Theoretical Impediments to machine Learning with seven sparks from the Causal Revolution. 3-3.

Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. New York: Basic Books

Pontin, J. (2018, February 2). Greedy, brittle, opaque, and shallow: the downsides to deep learning. Retrieved from https://www.wired.com/story/greedy-brittle-opaque-and-shallow-the-downsides-to-deep-learning/ on Nov. 27, 2018

Somers, J. (2017, September 29). Is AI riding a one-trick pony? Retrieved from https://www.technologyreview.com/s/608911/is-ai-riding-a-one-trick-pony/ on Nov. 30, 2018

Stanford Encyclopedia of Philosophy (2001). Counterfactual Theories of Causation
Retrieved from https://plato.stanford.edu/entries/causation-counterfactual/#Lew197CouAna on Nov. 26, 2018