The Alignment Problem: Machine Learning and Human Values

By Brian Christian. New York: W. W. Norton & Company; 2020. 496 Pages. $9.99

Reviewed by Declan Grabb, MD

DOI:10.29158/JAAPL.240036-24

Key words: artificial intelligence; risk; forensic psychiatry; technology

In the last year, artificial intelligence (AI) has newly entered our society's lexicon, being appended and prefixed to various words, terms, and tools that surround us. A topic that began in the 1950s in the recesses of computer science departments has now rapidly ascended to ubiquity.[1] AI and machine learning are having far-reaching effects in nearly every human pursuit: creative, academic, and recreational.[2–4] The respective disciplines that are being disrupted and changed by AI's influence are doing their best to adapt and restructure. Forensic psychiatry is not immune to such influence or disruption. Luckily, many individuals have already considered the broad-sweeping benefits and dangers of powerful AI technology, and there are decades of literature on this very topic.[5] In this review, I will highlight how the book *The Alignment Problem: Machine Learning and Human Values* does an exceptional job at summarizing AI's history while also focusing on various strategies to ensure machine learning and AI are "aligned" with human values. The text is the product of hundreds of conversations with experts in the field of artificial intelligence. The book is succinctly woven together by author Brian Christian and received the Excellence in Science Communications Award from the National Academies.

Christian writes, ". . . how to ensure that these [AI] models capture our norms and values, understand what we mean or intend, and, above all, do what we want — has emerged as one of the most central and most urgent scientific questions in the field of computer science. It has a name: *the alignment problem*" (Ref. 6, p 12). The book is divided into three sections: "Prophecy," "Agency," and "Normativity."

Prophecy

In this section, the author dedicates one chapter each to representation, fairness, and transparency in AI.

AI models are trained on real-world data, which contain representations of the biases and inequities of our society. "It's true by definition that there is always proportionately less data available about minorities" (p 29), which results in AI models that perpetuate bias and perform worse in these populations. The astute point is made that these AI models are not deployed in a vacuum ("modeling the world as it is is one thing" (p 49)), but Christian notes that utilizing a biased model is "*changing* the world, in ways large and small" (p 49). As his discussion moves into fairness, he comments on the age-old desire to replace the subjectivity of human judgment with something more calculated and objective, especially in the realm of criminal justice. "The idea that society can be made more consistent, more accurate, and more fair by replacing idiosyncratic human judgment with numerical models is hardly a new one. In fact, their use even in criminal justice is nearly a century old" (p 51). He describes the historical context for algorithmic decision-making in court systems by describing Illinois' 1927 attempt to improve prediction surrounding parole violation. He recounts relevant technological and political developments throughout the 1960s and 1970s, leading to the recent backlash regarding such algorithmic decision-making, citing the headline "When a Computer Program Keeps You in Jail" (p 57–58). Finally, he discusses the concept of transparency in AI decision-making. (This is particularly relevant for inscrutable neural networks that contain innumerable layers.) "Many are finding themselves uncomfortable with how little they know about what's actually going on inside those models" (p 87), he writes. Christian explains how this "black-box problem" is relevant in the field of medicine, as we rely on AI models to quantify risks of rehospitalization, cardiac disease, and more. As these complex AI models govern an increasingly large portion of our world, there has arisen a sub-field of "interpretability" that seeks to understand how models arrive at certain conclusions.

Agency

In this section, the author dedicates one chapter each to reinforcement, shaping, and curiosity.

Christian begins this section by discussing historical developments in cognitive neuroscience (from

Stein to Thorndike to Klopf and more), paying particular attention to reinforcement of behaviors: how we perceive successes and failures and how this dictates our subsequent behavior. "The reinforcement-learning framework began opening up a vista on the fundamental problem of learning and behavior that would develop into an entire field, and would direct the course of artificial intelligence research through our present decade" (p 133), writes Christian. He succinctly describes how individuals have sought to apply lessons learned in cognitive neuroscience to the field of artificial intelligence. This then moves into a discussion of B.F. Skinner and behaviorism, whereby the researcher was focused on the "shaping up of behavior by reinforcing crude approximations of the final topography instead of waiting for the complete response" (p 154). Essentially, to train anything to perform a complex task, it is prudent to reward steps along the way. This was a lesson that did indeed revolutionize the neuroscience of the time, but it also had a ripple effect in the field of AI. If you want to train a particular model to complete a complex task, perhaps you should consider "rewarding" it for small steps along the way. He astutely highlights that the previous two chapters have largely considered discipline and maximizing external rewards, yet they neglect the concept of initiative or intrinsic motivation. He then explores the successes of AI models that are built to contain a certain amount of intrinsic motivation, highlighting how these models can balance external reward and internal motivation.

Normativity

In this section, the author dedicates one chapter each to imitation, inference, and uncertainty. Using the example of self-driving cars, Christian discusses the importance of "imitation learning" (p 219) in teaching AI models complex tasks. For instance, he highlights that there are many values that must be weighed while driving a car: "We want to get from point A to point B as quickly as possible, though not by going over the speed limit—or, rather not by going *too* far over the speed limit, unless for some reason we have to—and by staying centered in our lane, unless there's a cyclist or a stopped car—and not passing cars on the right, unless it's safer to do so than not to do so . . . " (p 222) and so on. He emphasizes that it is difficult to "formalize" this into a list of concrete objectives; rather, we want to use "indirect normativity" where we can simply say, "Watch how I drive. Do it like this" (p 223). He discusses risks inherent in this form of learning, too. This discussion is enhanced by an analysis of how models may infer our values and motivations from our behavior. Finally, he emphasizes the uncertainty that still exists in the field and how it may affect other academic disciplines: "This motivates a number of questions—in medicine, in law, in machine learning—about just what impact is, how to measure it, and how our decision-making ought naturally to change as a result" (p 288).

Forensic Implications

AI has evolved beyond simple sentencing algorithms. AI models now have increasingly complex values that are shaped by their training, which guides their behavior. A solid background in AI, which this book provides, is important for forensic psychiatrists to be able to intelligently engage in discussion about its inherent risks.

Patients are increasingly interacting with AI models in mental health tools and everyday applications, and these AI models have increasing agency and power. This means soon an AI model may be able to operate your computer, switching between windows, clicking on links, and purchasing items for you.[7] It is vital, then, to support AI models that value human life, both the prevention of harm to oneself and others. We need to question how these increasingly agentic AI models are going to handle suicidality and homicidality.

AI models are trained on data that contain societal biases deeply set within them. Most of the time, model developers employ tools to identify and combat these biases in an attempt to "align" the model with human values. Even if models have undergone this process, they may still contain hidden or residual bias. A desire to quantify and objectify decision-making with a specific algorithm does not protect oneself from imbuing bias into any system. In the book, this was exemplified in biased sentencing algorithms that perpetuated biases instead of accurately assigning parole. Regardless of whether an AI model is used in treatment or sentencing, we need to know how to identify and monitor specific forms of bias in algorithms.

AI models are going to increasingly make predictions in care. They may recommend the medication that is chosen for a patient, or they may recommend involuntarily admitting an individual to prevent suicide. As referenced earlier, the nascent field of "interpretability" is rapidly growing and pertains to understanding how models arrive at their conclusions. It is important to consider how much concrete reasoning we will demand our models demonstrate to trust their recommendations, and what occurs when these recommendations cause harm.

If forensic psychiatrists do not choose to engage in this technical field, they are opting to be excluded from these very important discussions, an omission that we may come to regret.

References

1. Martinez D, Malyska N, Streilein B, *et al.* Artificial intelligence: Short history, present developments, and future outlook. Cambridge, MA: Massachusetts Institute of Technology; 2019. Available from: https://www.ll.mit.edu/sites/default/files/publication/doc/2021-03/Artificial%20Intelligence%20Short%20History%2C%20Present%20Developments%2C%20and%20Future%20Outlook%20-%20Final%20Report%20-%202021-03-16_0.pdf. Accessed May 9, 2024
2. Morrone M. I asked ChatGPT to be my life coach. The results were surprisingly helpful [Internet]. Fast Company; 2023 July 18. Available from: https://www.fastcompany.com/90923620/i-asked-chatgpt-to-be-my-life-coach. Accessed May 9, 2024
3. Zeng X, Wang F, Luo Y, *et al.* Deep generative molecular design reshapes drug discovery. Cell Rep Med. 2022 Dec; 3(12):100794
4. Lenharo M. Google AI has better bedside manner than human doctors—and makes better diagnoses. Nature. 2024 Jan; 625 (7996):643–4
5. van de Poel I. Embedding values in artificial intelligence (AI) systems. Minds & Machines. 2020 Sep; 30(3):385–409
6. Christian B. The Alignment Problem: Machine Learning and Human Values. New York: W. W. Norton & Company; 2020
7. Ghaffary S. Tech companies bet the world is ready for “AI agents.” Bloomberg [Internet]; 2024 Feb 15. Available from: https://www.bloomberg.com/news/newsletters/2024-02-15/tech-companies-bet-the-world-is-ready-for-ai-agents. Accessed May 9, 2024

Disclosures of financial or other potential conflicts of interest: None.