BOOK REVIEW

Artificial Intelligence: Harrowing or Hopeful Job Futures?

The Alignment Problem: Machine Learning and Human Values, by Brian Christian (NY: Atlantic Books, 2021)

Reviewed by Alexander Chestnut

How do we align the interests of artificial intelligence with our own? In confronting this question, Brian Christian relies almost entirely on the research and experiences of others, conveyed through conversations, to make his narrative accessible to the public. He splits it into nine distinct chapters: "Representation," "Fairness," "Transparency," "Reinforcement," "Shaping," "Curiosity," "Imitation," "Inference," and "Uncertainty."

In "Representation," Christian introduces the reader to image recognition technology, highlighting the problems posed by unrepresentative training data in unsupervised learning. In "Fairness," he demonstrates the predictive power of statistics and neural networks in matters of criminal justice, emphasizing both the potential dangers of misusing these tools and the problems created by these tools' capacity to alter the world they're designed to predict. In "Transparency," Christian highlights the danger of the 'black box' that neural networks tend to become while introducing both more transparent alternatives and various means to look inside that closed system

In "Reinforcement," he then introduces the reader to reinforcement learning, in which one trains a neural network using rewards directly connected to one's primary objective. He also draws connections on several occasions between this reward-based methodology and human psychology before re-presenting the alignment problem itself. The problem with reinforcement learning, as Christian presents in "Shaping," is its ineffectiveness in solving complex tasks with few opportunities to offer rewards. Although this can be solved through the division of these tasks into simple steps, either directly or through the creation of extra incentives, but such incentives need to be crafted with great care. Two such incentives, which Christian takes much time to explore in "Curiosity," are novelty and surprise, which proved surprisingly effective in controlled environments, even standing alone . Here, he also confronts the problem of medium maximization, in which an intelligence's pursuit of its given incentives causes it to deviate from its true objectives.

In "Imitation," Christian presents an alternative means of training, in which an intelligence is asked to mimic the actions of human agents. Although this methodology is helpful for tasks that are difficult to describe , it requires more direct human involvement in order to gain resilience, and the problems that occur without such interventions serve to highlight the dangers of imperfect imitation. Such imitation became an inspiration for the Inverse Reinforcement models discussed in "Inference," which would observe human behavior in an attempt to replicate the reward functions that create such behavior. This would enable agents to complete tasks that are beyond human capabilities as well as to cooperate with humans directly. However, this relies on the assumption that human actions advance their own interests. That human and artificial agents are fallible, as Christian notes in "Uncertainty," highlighting the need for both to be able to recognize situations where they don't know what to do. Similarly, it is imperative that powerful artificial agents understand the impacts of their actions. Although some provisional answers to these needs exist, they are hardly perfect.

In each of these chapters, the author seeks to highlight the progress made in answering the challenges posed by artificial intelligence while still leaving some unsolved problems lingering in the reader's mind, only to return the reader's attention to those problems and more in his conclusion. He further emphasizes the enormous present and potential capacity of artificial intelligence to reshape our society and our world, ultimately in furtherance of his claim. Mr. Christian argues that, although our models of the world and of ourselves may be improving, we are still not ready to entrust our fates to them, nor to the artificial intelligences that use and embody them.

Many of the artificial intelligences that Christian discusses in The Alignment Problem, especially in the first three chapters, can be described as models, designed to use training data supplied by humans to make predictions applicable to a general population. Such was the case with the Perceptron, a software designed to use images of shadows to identify the location of the source. This methodology, however, can run into serious problems. Such is the case with Word2Vec, a system designed to transform words into vectors, thus establishing relationships between those words in a manner similar to analogies. These seemed fairly logical, until the system's creators typed "doctor – man + woman" and received "nurse" as an output . Given that Word2Vec was designed only to create vectors for words based on how they were used in the training data's sentences, one eventually reaches the conclusion that the system's sexist outputs must be the result of sexist biases in the training data . A similar conclusion was reached in the case of Google Photos, after its image-recognition software misidentified two African Americans as "Gorillas." Here, however, the problem was the absence of images of people of non-white ethnicities in the training data, likely caused by systematic selection bias. Thankfully, in these cases, the instances of bias were mostly harmless, but that cannot hold true forever.

Such danger becomes significantly more visible in Rich Caruana's work designing a neural network for Pittsburgh hospitals to predict outcomes and prescribe care in pneumonia cases. Although Caruana's neural network showed the greatest predictive accuracy among the models tested, it did show some rather strange behavior. For instance, his software noted an association between asthma and positive health outcomes, which should be facially nonsensical but became rather clear upon conversation with some doctors. Generally, patients with asthma are more vulnerable than the general population, especially when they suffer additional respiratory problems, and the doctors know this too. Thus, pneumonia patients with asthma are almost immediately placed, not merely in the hospital, but in the intensive care unit, explaining their more broadly positive outcomes. Such a confounding variable clearly was not, nor could it have been, caught by Caruana's software, and thus relying upon it not merely as a predictive tool but as a prescriptive one would have proven disastrous upon implementation. Thankfully, it was not selected in this case, due in part to Rich Caruana's own protests against its use.

Even for agents that learn in more supervised ways, we must still contend with the fact that we need a model for our values. In reinforcement learning, this model takes the form of a set of rewards, given when autonomous agents perform actions that we describe. For simple tasks, rewarding the end-goal, be it points, profit, or victory in a game, is sufficient. More complex tasks, however, often demand shaping rewards designed to direct an AI toward our desired outcome . Actually achieving that goal, however, is easier said than done, especially when the outcome desired is more complex than merely scoring points, and designing the rewards incorrectly can lead to results that are undesirable to say the least. In the case of a knowledge-seeking agent, for instance, Brian Christian raises concerns that such an agent could end up "commandeering various earthly resources" in its pursuit if released into our world .

In a similar vein, Christian argues that agents reliant on imitating or cooperating with humans must create a model of human behavior that carries a few critical assumptions. The most important of these is expertise, under which the human both "knows what they want, and is doing ... the right things to obtain it." Otherwise, we risk sending an agent in the wrong direction and causing it to take undesirable courses of action. Christian finds an example of this in addiction, for an addict will desire and pursue the substance to which they are addicted while knowing that its consumption will harm them. An agent created through cooperative inverse reinforcement learning, though, will fail to recognize this, instead seeking to supply the addict with more of the substance in question, at best hindering their recovery and at worst causing further bodily harm.

Although his economic views in this work are somewhat difficult to ascertain, I have reason to believe that his arguments are most consistent with those of the Industrial Relations school of thought. This is due to the fact that the artificial intelligences that could serve us in the future are likely to have conflicts of interests between those of their owners and those of their creators , likely because their creators may be in pursuit of profits beyond the initial sale. This idea of a conflict of interests in a capitalist system may bleed into the labor market. What is telling here is that Christian's solution is not to eliminate one of the conflicting parties, as a Marxist would suggest, but rather to attempt to balance the interests of the parties by the creation of new protective legislation. This firmly places Christian's views within the Industrial Relations school of thought.

Although the vast majority of the evidence Brian Christian offers the reader is anecdotal, it is sufficient to demonstrate the enormous potential dangers of an unaligned artificial intelligence and of reliance upon flawed models. He further succeeds in highlighting the enormous difficulty, but not impossibility, of creating accurate models, not merely of the world, but of our values and of our behavior.

Alexander Chestnut is a recent graduate in Economics at Hofstra University.

REGIONAL LABOR REVIEW, vol. 26, no. 2 (Spring-Summer 2024).