Indian Journal of Medical Ethics Published online first on October 11, 2024

BOOK REVIEW

Can machine learning ever be taught to reflect the uncertainty and cultural relativity of human values?

JOHN H NOBLE JR

Brian Christian. *The Alignment Problem: Machine Learning and Human Values.* New York, NY: W.W. Norton & Company, 2020. 476 pages, ISBN: 9780393635829.

As stated by the author, "This book is about machine learning and human values: about systems that learn from data without being explicitly programmed, and about how exactly—and what exactly—we are trying to teach them."[p 11] The author interviewed many of the creators of machine learning, a.k.a. Artificial Intelligence (AI), and tells the story in their own words. Telling machines how to programme themselves is particularly challenging because so much of human communication is itself ambiguous. Listen into any conversation between two people. What do you hear back and forth? "Say it again," "What do you mean?", "I don't understand." Pity the poor machine trying to make sense of it when the humans have a hard time! No wonder there is a mismatch between machine learning and human values. Yet the creators succeeded to the point of AI becoming an everyday tool creating benefits and costs for its users. The US Congress and Office of the President are struggling to contain threats to its unbridled use while promoting beneficial applications [1].

The book is loosely organised and a challenge to read. There is no statement about methodology, but one that could be construed as such is found in the Acknowledgements: "This book is a product, more than anything, of conversations: many hundreds of them" [p 331]. There is a prologue, an introduction, nine chapters, a conclusion, acknowledgements, notes, bibliography, and an index. Yet the story is captivating and worth reading. It is a good place to begin learning what machine learning is and how it has become a forceful contemporary reality.

Author: **John H Noble, Jr** (jhnoblejr@icloud.com), Professor Emeritus, University of New York at Buffalo, New York, USA.

To cite: Noble JH Jr. Can machine learning ever be taught to reflect the uncertainty and cultural relativity of human values? *Indian J Med Ethics.* Published online first on October 11, 2024. DOI: 10.20529/IJME.2024.063

Manuscript Editor: Sanjay A Pai

The author was a science reporter with limited understanding of mathematics and the numerous disciplines that were ultimately blended to produce existing AI systems. Many of his informants created advanced mathematical models and consulted with experts in the physical and social sciences. There are many insights about how scientists operate and how science learns from conflicting theories and theoreticians — most importantly, those in the behavioural sciences. The principles of BF Skinner's operant conditioning theory and I Pavlov's alternate conditioning theory of associative learning have contributed to large segments of AI development.

Indeed, Pavlov's associative learning in the end points to Christian's ultimate view of where AI will contribute to the wellbeing of society, including enhancement of medical practice and ethics [p 124]. Associative learning is the foundation for reaching individual and civic self-knowledge. Current biased and unfair models — especially when linked to neoliberal purposes — endanger societal enhancement and advance. Given the *IJME* commitment to improved medical practice and ethics, I will focus on the contents of Christian's book that contribute to that goal.

Pavlov's associative learning is based on what Thorndike calls the "law of effect" — simply put, "connections leading to satisfying outcomes are strengthened while those leading to unsatisfying outcomes are weakened. Positive emotional responses, like rewards or praise, strengthen stimulus-response. Unpleasant responses weaken them."[2] Christian postulates unorganised machines "borrow directly from what was known about the nervous system, and the 'course of education' would borrow directly from what the behaviorists were discovering about how animals (and children) learned."[p 125] Christian documents "how difficult it is to create a reward function . . . that will engender the behavior you want, and not entail loopholes or side effects or unforeseen consequences." He characterises the belief of many in the AI field that handcrafting explicit reward functions as "a kind of well-intentioned road to hell, no matter how thoughtfully, . . . or how pure your motives." [p 300]

What is the best way to describe for physicians the benefits and costs of AI? The matter is not new; indeed, Plato describes what Socrates has to say: "Knowledge is a fine thing quite capable of ruling man; if he can distinguish good from evil, nothing will force him to act otherwise than as

knowledge dictates, since wisdom is all the *reinforcement* he needs" (my italics). Christian argues that every AI algorithm reveals a connection to ancient Greek philosophy, and, I might add, to ancient Indian philosophers, eg, Shankaracharya, the father of Indian philosophy [3].

Christian's book index lists 10 items for medical applications and medical predictive models [p 466]. For our purposes, the "uncertainty" link is a good starting point because medicine is generally acknowledged to be an art that depends on evolving physical, behavioural, and social science. The 1959 classic describing the dilemma is *Experiment Perilous: Physicians and Patients Facing the Unknown* by Renee Fox [4]. Diagnostic and treatment protocols are, at best, based on the mean effect of an unbiased clinical trial. But patients are individuals whose individual reactions to treatment vary within an estimated confidence interval of that mean. I have described the statistical and political issues involved elsewhere [5].

Christian stresses the observation of Yarin Gall, leader of the Oxford Applied and Theoretical Machine Learning Group, that teaching, "before any code is written or theorems proved or models trained, is almost entirely philosophy."[p 282] He goes on with the example of a physician using a model to diagnose if a patient has cancer and whether to start treatment or not, stating "I wouldn't rely on a model that couldn't tell me whether it's actually certain about its predictions."[p 283] There are dangers in relying on models that do not disclose whether they are certain about predictions. Bayesian neural networks may point the way to a solution because they explicitly encode a probability distribution over what range of numbers could be used to indicate what might be the output's certainty or lack thereof. The beauty of this solution is that the user can draw random samples from them to assure that the model doesn't give the same prediction every time. Alas, this doesn't solve the problem by itself. Instead, the user hits a computational wall. The beautiful mathematics of it is "of limited use for a long period of time when you want to do actual applications."[p 284] So, is there a solution to the problem, and what might it be?

The solution is to create an algorithm that quantifies and controls the uncertainty of a decision, allowing the physician user to "know when and whether she is uncertain about a case," and to consult with a human specialist if need be. The case example comes from a group at the Institute for Ophthalmic Research at the Eberhard Karls University in Tubingen Germany, led by Christian Leibig. The system they created knew what it didn't know about diagnosing diabetic retinopathy, a major cause of blindness in adults [6].

The interested physician or medical sociology reader can locate all 13 medical application and/or ethics references in Christian's Index [p 466].

Note: *The link to reference 5 was updated on October 22, 2024.*

References

1. Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence, Oct 30, 2023. 88 FR 75191
2. McLeod S. Edward Thorndike: The Law of Effect. *Simply Psychology.org* 2024 Feb 1[Cited 2024 Sep 20]. Available from: https://www.simplypsychology.org/edward-thorndike.html
3. Roy A. Indian Philosophy: Orthodox and Heterodox Schools. Rammohan College. [Cited 2024 Sep 20]. Available from: https://www.rammohancollege.ac.in/images/Study%20Materials/Indian%20Schools%20of%20Philosophy-Anima%20Roy.pdf
4. Fox R. *Experiment perilous: physicians and patients facing the unknown*. Glencoe IL: The Free Press, 1959.
5. Noble JH, Jr. Detecting bias in biomedical research: Looking at study design and published findings is not enough. *Monash Bioeth Rev.* 2007 Jan-Apr[cited 2024 Oct 22]; 26(1/2): 24-45. https://doi.org/10.1007/bf03351464. Available from: https://www.researchgate.net/profile/John-Noble-Jr/publication 5974020_Detecting_bias_in_biomedical_research_looking_at_study_design_and_published_findings_is_not_enough/
6. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. *Sci Rep*. 2017 Dec 19;7(1):17816. https://doi.org/10.1038/s41598-017-17876-z