Aaron Doerfler

Dr. Charles Luke Stark

MIT 4031G: AI, Ethics & Human Health

Friday, May 4th, 2021

Putting Flesh on the Bones of AI: A Literature Review

Technological progress over the last five years has placed humanity into a society full of applications of artificial intelligence and machine learning. These applications are incredible technological feats that can be applied to many sectors of modern life, including finance, healthcare, and law. However, recent work in artificial intelligence research questions the ethical implications of the current problems that surround the wider adoption of machine learning in various societal disciplines. Authors such as Trooper Sanders, Anneke G. van der Niet, Alan Bleakly, and Chen et. al consider the ethical implications of artificial intelligence considering potential regulations and pushing transparency in the technological development of these machine learning systems.

To begin, Trooper Sanders vows for more visibility and transparency in the development of artificial intelligence. He states that "institutional investors can use their influence to bring greater transparency to AI in two ways: pushing regulators to demand more disclosure by public companies and assessing the ethical AI fitness of companies in their portfolio who have materially significant stakes as AI developers or consumers" (Sanders 106). He uses the example of climate change to illustrate this idea, explaining that hard law and regulation forces companies to think about how the climate crisis affects businesses and investors, instead of waiting for the company to explore it on their own (106). Thus, Sanders feels that artificial intelligence should

use laws and regulations to ensure that there are ethical considerations when they are being developed. He quotes Virginia Dignum, an AI ethicist, who says that "engineers are those that ultimately will implement AI to meet ethical principles and human values, but it is policy makers, regulators and society in general that can set and enforce the purpose" (107). This suggests a need for diverse groups, that lie beyond the technical engineers who created the machine, to express ethical concerns about artificial intelligence. Sanders then summarizes his argument when he writes that "institutional investors' involvement in AI ethics is no balm to the havoc rogue AI can cause, they can be constructive allies in the push to align the power of technology and the public interest. Whether putting money behind ethical performance yields returns that sustain their interest depends on pressure from and decisions by developers, regulators, and consumers who drive AI's course" (108). This reveals the underlying theme of the piece that focuses on artificial intelligence being used in the public interest. This aligns with the topics of this course, where there was not a specific mention of healthcare, there are still heavy suggestions on how AI can incorporate ethical considerations into their modelling. The suggestion of transparency also relates to the "black box" problem in AI, where users can see the output, but not how the machine calculated that input.

In a similar fashion, Anneke G. van der Niet and Alan Bleakly consider some of the negative impacts of artificial intelligence, but with a specific focus on the healthcare industry. They argue that "machines lack human qualities such as empathy and compassion, and therefore, patients must perceive that consultations are being led by human doctors. Furthermore, patients cannot be expected to immediately trust AI; a technology shrouded by mistrust" (van der Niet and Bleakly 31). Instead, the authors emphasize a need to blend the technological innovation of AI with the more sensitive and personable human support that is reflexive of current medicine.

Specifically, “humans… are ‘open’ systems (complex, dynamic and adaptive) with multiple, competing feedback loops, and show many cognitive biases and emotional contaminations. But it is precisely in such colourations arising from uncertainty that medical consultations may gain meaning, or are given heart” (32). Thus, the authors suggest different strategies to ensure that machine learning devices do not lose the human side of medicine. Similar to the point raised by Sanders, the authors write that “despite the accuracy and efficiency of the performance (a gift box), AI is often a black box. This black box needs to be opened. If not, AI can place doctors in all kinds of difficult situations, providing ‘solutions’ that they do not understand or agree with, or that are erroneous, again raising ethical questions concerning responsibility and agency,” (34) and adding that “medicine has an intrinsic artistry and humanity that by definition is plastic and cannot be replicated by AI” (35). Just like Sanders, van der Niet and Bleakley emphasize that technological solutionism is not going to solve the ethical issues in artificial intelligence. They focus more on this issue in relation to healthcare but do a good job of illustrating the isolating feeling that an algorithm can impose in an industry with a large emotional component to it. This article is more specific to the context of this course since we are specifically speaking of the healthcare industry. Once again, the black box problem was highlighted, which we often talk about when it comes to data transparency and ethical considerations in preventing inequality in outputs.

Comparatively, Chen et al. provide a specific example of machine learning displaying bias and inequality in healthcare. In their study, the authors examine bias by measuring differences “in model error rates in patient outcomes between groups, and show that in the ICU data set, differences in error rates in mortality for gender and insurance type are statistically significant and that in the psychiatric data set, only the difference in error rates in 30-day

readmission for insurance type is statistically significant" (Chen et al. 169). The authors highlight that because machine learning models are created using large quantities of data, bias can be encoded into the modelling choices or even within the data itself (168). To be more specific, the authors found that "for gender, female patients have a higher model error rate than male patients; for insurance type, public insurance patients have a much higher model error rate than private insurance patients," (171), and further, "black patients having the highest error rate for psychiatric readmission" (173). Since these disparities do exist, the authors seem to write in accordance with van der Niet and Bleakley, suggesting that a "closely cooperative relationship between clinicians and AI— rather than a competitive one—is necessary for illuminating areas of disparate health care impact… Indeed, algorithmic scrutiny is vital to both the short-term and long-term robustness of the health care system" (174). This study represents a multitude of topics from the course including bias and inequality, data transparency, and looking into ethical considerations. It focuses on a specific study of artificial intelligence and healthcare which can be analyzed critically using the knowledge and topics the course has discussed.

Therefore, these authors are all in agreeance that there needs to be some form of intervention or change in the way that machine learning devices are made. However, there is disagreement in the suggestions for these changes. For example, Sanders argues the importance of regulation whereas van der Niet and Bleakley are focused on the need for more human presence in AI, specifically towards healthcare. This is a good representation of the current state of research surrounding ethical considerations of machine learning. It illustrates the idea that these are complex problems, and there are no easy concrete solutions that have been determined to completely change AI for the better. However, one issue that all authors agreed on was the need to solve the black box problem, which is ensuring that creators of machine learning

applications are being transparent in how their algorithms are reaching the conclusions that they are. This shows that these articles agree that there are many biases currently in artificial intelligence. The authors highlight different forms, for example, “ethical questions [surrounding] this kind of application [including] data protection, third party intervention and poor coding (including algorithmic bias towards white ethnic groups),” (van der Niet and Bleakley 33) of algorithmic bias that ethical considerations need to tackle. This was different depending on the article, which is overall good for scholarship in general. If more problems are highlighted, more solutions can be applied to them, incorporating ethical considerations into machine learning.

Thus, I conclude that recent work in artificial intelligence research questions the ethical implications of the current problems that surround the wider adoption of machine learning. This is illustrated by each author, where Sanders poses solutions such as regulations and hard-law to van der Niet and Bleakley thinking more about transparency through solving the black-box problem and bringing a more human presence to healthcare. The article from Chen et al. illustrates the current biases in artificial intelligence, citing specific issues with women and people of colour in intensive care units, hence highlighting why scholarship is focused on correcting the current issues surrounding artificial intelligence in various societal disciplines. These articles are a good representation of the diverse thought process in the current scholarship that surrounds artificial intelligence and machine learning. Whereas technologists are concerned with creating the best product, scholars are concerned with “[putting] flesh on the bones of AI,” (van der Niet and Bleakley 35) helping to further incorporate machine learning processes safely into everyday life.

Works Cited

Chen, Irene Y., et al. “Can Ai Help Reduce Disparities in General Medical and Mental Health Care?” *AMA Journal of Ethics*, vol. 21, no. 2, 1 Feb. 2019, https://doi.org/10.1001/amajethics.2019.167.

Sanders, Trooper. “Testing the Black Box: Institutional Investors, Risk Disclosure, and Ethical AI.” *Philosophy & Technology*, vol. 34, no. S1, 24 July 2020, pp. 105–109., https://doi.org/10.1007/s13347-020-00409-4.

Van der Niet, Anneke G., and Alan Bleakley. “Where Medical Education Meets Artificial Intelligence: ‘Does Technology Care?’” *Medical Education*, vol. 55, no. 1, 30 Feb. 2020, pp. 30–36., https://doi.org/10.1111/medu.14131.