Morality, Capability, and the Ethics of Decision-Making AI

Andrew E. Lawrence
Tufts University

December 15th, 2023

Addendum:
This paper was originally written for my second-year course, STS-0057: Philosophy of Technology, with Jamee Elder.

Research Question

Do decision-making algorithms perpetuate discriminatory practices and exhibit unfair or unjust behaviors[1]? If so, what are the causes and is it morally permissible to employ these systems in our daily lives?

Introduction

Published circa 300 BC, Euclid's *Elements* introduces a series of steps to find the greatest common divisor of two integers, *a* and *b*. This algorithm, now known as the Euclidean Algorithm, has varied minimally over the millennia and is still used by programmers today—even with its unremarkable computation complexity of O(n) ("Euclidean algorithm"). Yet, with a quick glance, one can see its scope of impact differs vastly from those developed today. Rather than helping a mathematician solve a simple problem, increasingly algorithms are tasked to interpret complex scenarios that affect the decision-making processes of those implementing and engaging with them.

This presents an issue as, unlike Euclid's Algorithm, which has stood largely unchanged over millennia due to its logically consistent results, modern decision-making programs struggle to produce unquestionable outcomes. In fact, research suggests that many systems actively discriminate or are unjust during their computations. Consequently the widespread implementation of these algorithms begs serious ethical analysis. This paper will first demonstrate that software does act unfairly, then explain why these supposedly unbiased algorithms do, in fact, discriminate, then finally argue that given our current computing capabilities and the basic nature of our concepts of fairness and justice, we should not be attempting to impose these judgements onto algorithms.

Background

Unlike Euclid's days, algorithms—which by definition are simply a series of rules to follow to compute some result—are incredibly complex. Furthermore, they are playing an increasingly important role in our daily lives: Google Maps directs us through traffic, we scroll through Instagram all day, and Indeed chooses what jobs to display—all through an almost uncountable

[1] Note that I will be using the terms "fair" and "just" interchangeably throughout this paper. While they are distinct philosophical terms, they are closely tied and the variation is necessary for readability.

number of steps that produce the outcome we look towards.

Yet, these seemingly systematic and unbiased systems are in fact capable of reinforcing widespread discrimination. In 2000, Northpointe, Inc. introduced a comprehensive software package to predict recidivism called COMPAS. Hailed as an effective way to provide a neutral assessment of a criminal's likelihood to recommit crimes, it was implemented in many states across the US (Angwin). However, actual use cases of COMPAS suggest it did not. In 2014, COMPAS assigned a lower score of "3" to a repeat white offender, who had previously been convicted of armed robbery twice, than to a 18 year old black girl, who had only ever committed petty theft, and received an "8" (Angwin). While this outcome seems incorrect, developers at Northpointe and elsewhere argued that, mathematically, their software gave the same probability of recidivism for each race and thus the results were accurate and without prejudice (Green). Yet overarching empirical evidence suggests differently. Black defendants were twice as likely than white defendants to be falsely labeled as future criminals, and white defendants were falsely labeled as low risk more often than black defendants (Angwin). In fact, according to a 2018 study by Dartmouth Professors Julia Dressel and Hany Farid, "COMPAS's collection of 137 features" to determine recidivism had the same accuracy as a "simple linear classifier with only two" (Dressel). Clearly this algorithm, which has been implemented across our country, has some serious controversy and some likely deficiencies. Given COMPAS's recidivism rating influences a judge's decision on the severity of punishment a defendant deserves, COMPAS's continued use across the US is undeniably problematic. In fact, with a quick glance, one could claim that COMPAS's incredible mistreatment of black suspects violates the US' 14th amendment on unequal treatment based on race. However, COMPAS is not the only software to exhibit seriously discriminatory practices. In a 2023 analysis of Large Language Models (LLMs), massively complex algorithms that predict and generate text, Apple researchers noted that LLMs contain numerous gender biases and stereotypes encoded into their processing. Notably, when prompted to align a job to a person, these programs were "3-6 times more likely" to assign a corresponding stereotypically gendered role to that person, regardless of the actual occupation gender statistics (Kotek). Furthermore, they are unable to give accurate reasoning for these assessments and instead obscure the rationale for their predictions (Kotek). This analysis would suggest that LLMs, which are being increasingly used in the workforce to automate tasks, exhibit prejudices. This lies in direct contrast to the sentiments of many who take the responses of LLMs like ChatGPT or Google Bard to always be accurate. Whether or not the average person utilizes some piece of software knowing the potentiality for it to be flawed, it is clear that such programs—which increasingly play greater roles in our lives—cannot provide the consistent and fair results that many developers suggest they do.

Main Argument

The two examples provided above are among the countless software platforms that cannot define or provide a consistently just result. However, while empirical evidence highlights this issue, in order to further understand these challenges, it is important to look at the methodological background which causes them.

Currently, software engineers attempt to develop platforms that are considered "fair" through mathematical notions of justice which are then integrated into their systems. However, developers are unable to define and implement fairness as a mathematical concept useful to programs, creating an enormous developmental roadblock which obstructs progress.

Although there already exist various ways to calculate justice, not only is it impossible for algorithms to be considered fair under multiple equations, but there are also numerous issues regarding how fairness is currently defined. Similarly to the construction of important philosophies, one outcome an algorithm gives could be considered fair under one principle and unfair under another. Likewise, fundamental differences in what is considered moral, or in this case fair, cause an algorithm to be unable to satisfy the requirements for multiple arbiters of fairness. In "Algorithmic Fairness and Statistical Discrimination," John Patty, professor of political science and quantitative theory at Emory University, breaks down the fundamental differences between the two popular ways of determining a fairness value, Predictive Parity and Error Rate Balancing, that cause this problem. Predictive Parity rates the fairness of an algorithm based on the "applicants' outcomes conditional on the algorithm's decision" (Patty). However, Error Rate Balance instead determines the algorithmic fairness (AF) given the "algorithm's decisions conditional on the applicants' outcomes" (Patty). This creates a base difference that will come into conflict. To illustrate this potential, Patty examines how the two forms of determining fairness would contradict in a hiring scenario. Suppose some algorithm used information about a series of insensitive and sensitive traits - qualities that are acceptable to differentiate on, like test scores, or not, like gender - to determine which applicants would obtain some job. Then, Predictive Parity would necessitate that the outcomes among those hired are distributed equally across various sensitive traits, where Error Rate Balancing would cause the algorithm to focus "on fairness conditional upon individual outcome" - altering an applicants' decisions to apply at all (Patty). This distinct difference causes decision-making systems to be unable to accommodate both justice indicators. However, it does not seem that satisfying only one is enough to be adequately fair. In a 2022 analysis of AF, Ben Green, an assistant professor at the University of Michigan School of Information, suggested that the AF indices themselves "'[mirror] some of antidiscrimination discourse's most problematic tendencies' as a mechanism for achieving equality," which causes a "significant gap" between evaluations of fairness and real-world impact that "[exacerbates] oppression and [legitimizes]

unjust institutions" (Green). This analysis is consistent with the two examples of algorithms perpetuating inequality given above, further highlighting that the current methodologies of promoting AF fall short of their goal. Given this, it would seem that we should not continue to employ these algorithms. Specifically, if we were to analyze their usage from behind the social contract theory, we would see that the widespread discrimination these algorithms are capable of creates an unjust environment where the disadvantaged—those who the system draws conclusions from—cannot escape discrimination. As a result, these programs cannot be rationally justified and thus they should not be utilized.

However, even if current approaches to quantifying and applying justice in decision-making algorithms were flawless, many tech philosophers question the base definition and methodology for determining fairness. In his same analysis of algorithmic fairness, Ben Green extends his claim that AF is currently impossible. He goes on to suggest that "formal algorithmic fairness" restricts analysis to "isolated decision-making procedures" and decision points that are narrowly focused and thus flawed (Green). Not only is this overtly problematic because these processes are impacting the lives of millions, but it would also suggest that the defining feature of algorithms—the separation of steps into smaller functions—enables them to perpetuate discrimination. This is a concerning conclusion to draw, yet Ben Green is not the only one to come to it. In "A sociotechnical view of algorithmic fairness", Mateusz Dolata et al. contend that fairness is inherently a social construct and thus the present techniques for determining what is just are inadequate. In fact, they suggest that attempting to determine justice within an algorithm is in of itself problematic as "achieving fairness remains an ongoing process rather than a one-time challenge" (Dolata). Such is the crux of the problem.

The distribution of "fair" software relies on a general consensus of opinion on what is fair. Even if some algorithm were to obtain equally accurate predictions across groups, generally referred to as an equality of accuracy, it might still be considered unfair in practice because it does not meet predictive parity—the measure of fairness discussed earlier. Furthermore, even if it did, fairness is an ever changing notion. Today's software engineers might develop the perfect algorithm just for tomorrow's public opinion to decide it acts unfairly. As a result, it seems that the existing approaches to integrating concepts of fairness into programs fall seriously short.

Considering the evident challenges AF faces, it begs the question whether these programs should be attempting to be fair in the first place. I argue they should not. Given the ever-changing and social nature of concepts of fairness and justice, systems that look to fully capture them in a steady state are molesting the essence of these principles. If it were feasible to continuously undertake a deep analysis of what our society thinks is fair, then it might be possible to deconstruct that notion and integrate it into our systems. However, at the moment this seems incredibly unlikely. As a result, it would seem that attempts to create

just algorithms only provide a false cover by which those who employ these systems can turn to when questioned about the integrity and impact of them. Considering this, I instead contend that the breadth of the usage cases for algorithms needs to be decreased. When Euclid developed his algorithm, he was working with purely mathematical concepts. However we are not. Increasingly, businesses attempt to automate processes in order to increase efficiency, causing progressively more complex decisions to be made by programs that are clearly flawed. While people are flawed too, at least we have the capacity for dynamic change reflective of our surroundings.

Objections & Replies

Ultimately there are two main arguments against a reduction of the decision-making capacity of algorithms: productivity and the capacity of artificial intelligence to learn. Primarily, the effect decision-making algorithms make on increasing efficiency is undeniable. For example, Metroplex Health System, a mid-sized primary healthcare provider in Texas, documented their switch to an algorithmic hiring process called Pegged. Their CEO, Carlyle Walton, contends that after switching they saw decreases in turnover rates and increases in quality scores ("Challenges and Benefits of Hiring Algorithms."). His story is not unique. Across the board, it seems that automating certain tasks raises productivity and profits. However, this does not change the possibility for the systems they use to discriminate or act unjustly in other ways. Rather it simply highlights another known fact: generating profits is still the highest priority. On the other hand, recent developments in the proficiency of artificial intelligence softwares could mean that algorithms are able to adapt to new situations - potentially providing the previously unfound adaptability needed to create a truly fair system. However, this is not fully accurate for two reasons. Primarily, as the example of gender bias in LLMs discovered by Apple researchers above demonstrates, AI output remains problematic. However, at a more fundamental level, current artificial intelligence is created through the combined output of thousands of algorithms that themselves are not changeable. Thus they define and categorize input with a given set of standards that ultimately determine what is fair. As a result, the adaptability of AI is really only surface level. Given this, it would seem that we are still unable to replicate the continuously shifting nature of what we find just on machinery.

Conclusion

From Euclid to Ada Lovelace, great minds have helped develop ingenious algorithms to simplify and automate complex tasks. Yet our modern programs do not just print outputs for us to interpret, they decide our next actions. If these decision-making algorithms were able to provide consistently unquestionable results, then no one could deny their value. Yet they do not. Instead, deep methodological choices enable them to reinforce status quo disparities and actively discriminate. Only by reducing the role they play in our lives

can we restore the power to decide what's best for ourselves.

Bibliography

Angwin, Julia, et al. "Machine Bias." ProPublica, ProPublica, 23 May 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 18 December 2023.

Britannica, The Editors of Encyclopaedia. "Euclidean algorithm". Encyclopedia Britannica, 27 Nov. 2023, https://www.britannica.com/science/Euclidean-algorithm. Accessed 18 December 2023.

"Challenges and Benefits of Hiring Algorithms." *Monster.com*, hiring.monster.com/resources/recruiting-strategies/workforce-planning/recruiting-algorithms/. Accessed 18 December 2023.

Dolata, Mateusz, et al. 'A Sociotechnical View of Algorithmic Fairness'. Information Systems Journal, vol. 32, no. 4, 2022, pp. 754–818, https://doi.org10.1111/isj.12370.

Dressel, Julia, and Hany Farid. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances*, vol. 4, no. 1, Jan. 2018, advances.sciencemag.org/content/advances/4/1/eaao5580.full.pdf, https://doi.org/10.1126/sciadv.aao5580.

Giovanola, Benedetta, and Simona Tiribelli. 'Weapons of Moral Construction? On the Value of Fairness in Algorithmic Decision-Making'. Ethics and Information Technology, vol. 24, no. 1, Jan. 2022, p. 3, https://doi.org10.1007/s10676-022-09622-5.

Green, Ben. 'Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness'. Philosophy & Technology, vol. 35, no. 4, Oct. 2022, p. 90, https://doi.org10.1007/s13347-022-00584-6.

Kotek, Hadas, et al. "Gender Bias and Stereotypes in Large Language Models". *Proceedings of The ACM Collective Intelligence Conference*, ACM, 2023, https://doi.org/10.48550/arXiv.2308.14921.

Patty, John W., and Elizabeth Maggie Penn. 'Algorithmic Fairness and Statistical Discrimination'. arXiv [Econ.TH], 2022, http://arxiv.org/abs/2208.08341. arXiv.