ST.GALLEN SYMPOSIUM

#collaborativeadvantage

THOUGHT PROVOKING IDEAS OF THE GLOBAL ESSAY COMPETITION 2022

Together with AI: Envisioning a More Equitable & Fair Future

Savina Kim is is one of the top 25 contributors to this year's Global Essay Competition Award. She studies at London School of Economics and attended the 51st St. Gallen Symposium as a Leader of Tomorrow.

Introduction

Artificial intelligence (AI) algorithms are increasingly being used in high-stakes decision making from employment, parole, credit lending to medical diagnosis. However, despite its efficiency, cost savings and relative "objectiveness," its influence in domains with life-long impact has raised a host of ethical, social and political concerns. Central among these is whether these algorithms can systemically reinforce discriminatory practices against historically underrepresented or disadvantaged members of society and thereby exacerbate societal inequities by delivering unfair outcomes.

Society today is no stranger to concepts of discrimination and historical injustices which have defined and restricted the opportunities of women, ethnic minorities, the young and many others. While we hope these are unlucky or regretful elements of the past, they have reemerged in new form alongside the tidal wave fueled by AI (Corbett-Davies et al., 2018; Greene, Hoffmann, & Stark, 2019; Hu & Chen, 2020). There is no shortage of evidence highlighting the existence of bias and discrimination in machine learning (ML) systems with African American and Latino borrowers being charged higher interest rates than white borrowers (Bartlett, Morse, Stanton, & Wallace, 2022), facial recognition software with higher error rates for darker-skinned individuals (Buolamwini, 2018), an automated hiring system recommending applicants of particular race, age or gender (Broek, Sergeeva, & Huysman, 2019; Raghavan, Barocas, Kleinberg, & Levy, 2020), a policing algorithm claiming that "black people reoffend more," (Berk et al., 2017; Ensign et al., 2018), or the infamous case of Apple's "sexist" credit card (Neil Vigdor, 2019).

The greatest burden of these known (and yet unknown) risks of AI is being borne by younger generations. As they begin adulthood, enter the work force in entry-level positions or apply for their first mortgage, they are experiencing this technological shift first-hand via algorithmic hiring decisions, credit approvals, digital identity and college admissions, for instance. On the other hand, older generations have significant influence over the next steps, trajectory and ultimate freedom of those who deploy these technologies as regulators, politicians and senior-level managers in industry. This poses a disconnect between those who are in the position to forecast and direct its future use and those whose lives will be most impacted long-term. Therefore, intergenerational collaboration is critical during this formative stage to establish a strategy which can guarantee a safer, fairer future.

In the following, I present a solution to the question: How do we make sure that AI is appropriately guided to create a more opportunistic and fair society which represents a state of the world we want to see in the future? This is presented as a five-step strategy which reflects three major themes: equity, the consideration of differential impacts on vulnerable subpopulations; democratic deliberation, the participation of all relevant stakeholders; and sustainability, the recognition of both short- and long-term effects as well as potential feedback loops of a system.

Step 1. Shift the perspective – It's not only a technical issue

An algorithm functions by maximizing a given objective; for example, current social media trends are fueled by maximizing clickthrough, or engagement. In other words, they are trained to learn what users want to see to satisfy their perceived interests; in reality, the algorithm is not learning what people want but instead learning to influence and modify users to become more predictable via quantifiable metrics (e.g., click behavior) in order to maximize reward over time. Consequently, these simple, yet powerful algorithms have "manipulated" users to extremes in their consumption veiled under optimization for the benefit (i.e., profit) of their supplier. This lesson can be similarly applied to biased AI. The algorithm itself is not "racist," for instance, but rather a consequence of an incorrect (or limited) objective, where maximizing accuracy alone is futile if the underlying data is the source of bias. This highlights the need to consider fairness as a core performance metric as opposed to a bonus or by-product.

While fairness is a universally desirable quality in society, it poses a challenge to achieve in practice. Scholars have proposed over 20 definitions for fairness, including demographic parity, disparate impact, equality of odds, and calibration (Caton & Haas, 2020; Mehrabi et al., 2019; Verma & Rubin, 2018); however, it is mathematically impossible to satisfy multiple definitions simultaneously and consensus is still lacking. While these efforts are commendable, without a framework that accounts for the social, political and historical contexts of the environment in which the system is embedded, generalpurpose formulas are insufficient as they neglect deeper, non-observational origination of harms. Therefore, operationalizing fairness requires a holistic approach which combines procedural and statistical considerations incorporated with justice concerns, cultural values and a moral position. The following discusses these elements in greater detail with real-world scenarios demonstrating its complexity and proposes solutions for rectifying this dilemma. First, we must answer – what type of fairness do we want?

Step 2. Select a world view – Sourcing the bias

A fundamental component of any statistical measure is base rate determination. However, this can depend on how we perceive the source of bias and the question: Do we want our data and influenced beliefs to reflect the current state of the world or its idealistic state? For example, Google image search results significantly underrepresent women as chief executives with only 10% of images in the top 100 appearing as women

(they make up 28% of these roles in reality (Lam, Broderick, & Hughes, 2018). The skewed representation not only impacts searchers' perception of women in society but can influence career choices of future generations. This reflects two opposing worldviews: (1) "What you see is what you get" which 3 assumes absence of structural bias in the data therefore variation is attributed to warranted deviations in base rates and (2) "We're all equal" which presupposes equal base rates therefore any observed deviations are a result of unwanted structural biases caused or shaped by society which should be corrected (Friedler, Scheidegger, & Venkatasubramanian, 2021).

To exemplify how these concepts impact fairness decisions, imagine a college admissions scenario. Different base rates in standardized exam results across ethnic groups could be a result of unequal opportunities; therefore, if our fairness objective were to fix this historically influenced, social injustice, selecting equal base rates would be preferred (e.g., quotas). However, in medicine where 99% of breast cancer occurs in women, a diagnosis model should acknowledge this discrepancy and would cause further harm if it enforced equal base rates across genders. This shows how statistical metrics alone can easily backfire without acknowledging the wider system in which these models are imbedded. Along with domainspecific considerations, to properly identify instances of wrongful discrimination and select relevant fairness tests, stakeholder consideration is also critical.

Step 3. Gather the stakeholders – Aggregating priorities and trade-offs

Defining fairness is often convoluted by ideological or cultural differences as well as competing priorities between involved or impacted stakeholders such as data scientists, industry experts, regulators and consumers or subjects (Hutchinson & Mitchell, 2018; Mehrabi et al., 2019). For example, when assessing lending risk, a bank may prioritize profit, consumers may prioritize individual fairness and application processing speed and regulators may prioritize accuracy and group fairness. Simultaneously, there are multiple layers of impact where a default event not only diminishes a bank's profit but can also worsen the financial situation of a borrower with a credit score decrease and vice versa. How do we ensure all priorities are considered?

This requires public, democratic deliberation and consensus amongst multiple stakeholders to determine what "society" thinks as fair and cannot be left solely to the responsibility of computer scientists to select their own desired notion of fairness (they too are biased). Tools such as the Ethical Matrix can help build a shared language around describing fairness risks and ensuring a delicate balance across several societal goals (O'Neil & Gunn, 2020). By taking the following steps: (1) gathering relevant parties, (2) identifying sensitive subgroups, (3) weighing diverse viewpoints, and (4) setting thresholds for select priorities, can we then aggregate them into code as quantifiable objective(s). Determining these thresholds requires understanding decision costs, which are discussed next.

Step 4. Weigh the costs – Measuring impact

Algorithms now have more control over what human beings see, read and learn than under any dictator in history. However, the effects we observe today were not predicted or "planned" at inception and thus programmed with minimal regulation. Therefore, to understand impact, one must undergo the arduous task of measuring the consequences fairness constraints can have on classification outcomes relative to 4 other trade-offs. In other words, what is the cost (or utility) of a positive or negative outcome to different subjects, including marginalized groups? The term "cost" can be interpreted as the broader impact of a model's classification error; beyond economic costs, this can also include social and opportunistic harms. For example, college admission has different utilities to applicants depending on their other acceptances. Decision outcomes can also affect those who are not individual subjects such as communities when concerning

incarceration. The following walks through real-world examples where decisions are made based on select normative distribution principles highlighting varying levels of impact.

An egalitarian may view the population as either intended recipients or victimized groups (Leben, 2020); for example, predictive policing tools use statistical measures to “go where the crime is." Here the purpose of a risk assessment tool is to accurately classify crime and any unfair error rates are merely sideeffects or collateral damage. However, this can result in damaging systemic racism through overenforcement of neighborhood “hotspots” via racial redlining and can perpetuate a self-reinforcing feedback loop where prediction becomes a self-fulfilling prophecy (Raji et al., 2020). Opposingly, a compensationbased approach for distributive justice may focus on those most negatively impacted. For example, the auditing toolkit Aequitas differentiates assistive versus punitive interventions, claiming the former be evaluated on equality of false negative (FN) rates, or failing to give rewards to those who deserve them, and the latter on false positive (FP) rates, or imposing punishment on those who do not deserve them (Saleiro et al., 2018). In credit eligibility algorithms, a consequentialist, utilitarian may claim the negative utility of FPs as significantly greater than of FNs due to investment loss. In criminal justice, the opposite by lowering the potential of more violence assuming that the distress caused by keeping an innocent individual in jail (FP) is less than allowing a dangerous prisoner go free (Berk et al., 2017). However, one could argue that FPs hold a higher social cost for peaceful black prisoners than peaceful white prisoners given they are often disproportionately subject to more severe treatment (ProPublica, 2016), a product of centuries-long discrimination including segregation, police brutality and underfunding of social resources. Therefore, this may suffer the added cost of perpetuating the cycle of deprivation in the black community.

These examples show that understanding differential impact can help select fairness metrics which consider the social costs and benefits of everyone. Unfortunately, selecting one approach often results in the detriment of others; however, that is the nature of moral choice and to responsibly mitigate these harms to our best ability is to develop a comprehensive, iterative response to them.

Step 5. Calibrate and repeat – Feedback into the future

Throughout this paper we have quickly realized that there is no easy or right answer to the fairness problem. Not because of computational or statistical hurdles but because fairness is a continuous process of social negotiation and weighing of competing values (Kleinberg, Mullainathan, & Raghavan, 2017; Raghavan et al., 2020). Therefore, it is important to consider AI not as a technical system, but as a sociotechnical system which dynamically affects and feedbacks into its environment. For example, if the bar for college admission is lowered for a subgroup, this may increase their average qualifications over time due to (1) a larger proportion of the next generation growing up in households with college-educated parents 5 and expanded opportunities and (2) the possibility of college incentivizing individuals to prepare academically (Chouldechova & Roth, 2018; Liu, Dean, Rolf, Simchowitz, & Hardt, 2018).

These long-term, intergenerational effects cannot be quantified via statistical or individual definitions of fairness alone but must consider the dynamic nature of a system’s impact on society and evolving justice concerns. One solution is to conduct subsequent analyses, as opposed to static learning, which considers compounding effects of model interactions, or a frequent monitoring procedure revealing cumulative effects of previously deployed systems and fairness interventions(Bakalar et al., 2021). A multidimensional, iterative method acknowledges that there is no silver bullet but encourages proactive analysis throughout the development cycle and flexibility to reassess post-deployment for unintended harms.

Conclusion

The AI revolution has brought forth a resurgence of a new kind of fairness risk where proper remedies have yet to be found and for which consensus does not yet exist. In response, this paper has proposed a five-step strategy for a more holistic approach alongside a discussion of outstanding challenges and opportunities as a foundation for future work. Despite its novel risks, AI also provides us a blank slate where we have a unique opportunity to rewrite the social contract, reassemble the distribution of resources in a more equitable manner and amend past errors by expanding opportunities for those disadvantaged at the starting line. Different generations, representing those who affect and will be affected by these technologies, are encouraged to collaborate and accept this mantle of responsibility to help rebuild society for a more fair, inclusive future for all.

References

Bakalar, C., Barreto, R., Bergman, S., Bogen, M., Chern, B., Corbett-Davies, S., ... Facebook, J. Z. (2021). Fairness on the ground: Applying algorithmic fairness approaches to production systems.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143, 30–56.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods and Research, 50, 3–44.

Broek, E. van den, Sergeeva, A., & Huysman, M. (2019). Hiring algorithms: An ethnography of fairness in practice. ICIS 2019 Proceedings.

Buolamwini, J. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1–15.

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning.

Corbett-Davies, S., Goel, S., Chohlas-Wood, A., Chouldechova, A., Feller, A., Huq, A., ... Shroff, R. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning.

Ensign, D., Frielder, S. A., Neville, S., Scheidegger, C., Venkatasubramanian, S., Mohri, M., & Sridharan, K. (2018). Decision making with limited feedback: Error bounds for predictive policing and recidivism prediction. Proceedings of Machine Learning Research, 83, 1–9.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im)possibility of fairness. Communications of the ACM, 64, 136–143.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. Proceedings of the 52nd Hawaii International Conference on System Sciences.

Hu, L., & Chen, Y. (2020). Fair classification and social welfare. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 535–545.

Hutchinson, B., & Mitchell, M. (2018). 50 years of test (un)fairness: Lessons for machine learning. FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 49–58.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Leibniz International Proceedings in Informatics, LIPIcs, 67.

Lam, O., Broderick, B., & Hughes, A. (2018). Gender and jobs in online image searches. Retrieved from https://www.pewresearch.org/social-trends/2018/12/17/gender-and-jobs-in-online-image-searches/

Leben, D. (2020). Normative principles for evaluating fairness in machine learning. AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 86–92.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. IJCAI International Joint Conference on Artificial Intelligence, 2019-August, 6196–6200.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54. 7

Neil Vigdor. (2019). Apple card investigated after gender discrimination complaints. Retrieved January 11, 2022, from The New York Times website: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

O'Neil, C., & Gunn, H. (2020). Near-term artificial intelligence and the ethical matrix. Ethics of Artificial Intelligence, 237–270.

ProPublica. (2016, May 23). Machine bias. Retrieved January 12, 2022, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 469–481.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7, 145–151.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., ... Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. IEEE/ACM International Workshop on Software Fairness, 18.