Long-Term Impacts of Fair Machine Learning

Machine learning models used to inform decisions create fairness concerns. This study assesses how people react to machine-aided decisions.

By Xueru Zhang, Mohammad Mahdi Khalili, & Mingyan Liu

FEATURE AT A GLANCE: Machine learning models developed from real-world data can inherit potential, preexisting bias in the dataset. When these models are used to inform decisions involving human beings, fairness concerns inevitably arise. Imposing certain fairness constraints in the training of models can be effective only if appropriate criteria are applied. However, a fairness criterion can be defined/assessed only when the interaction between the decisions and the underlying population is well understood. We introduce two feedback models describing how people react when receiving machine-aided decisions and illustrate that some commonly used fairness criteria can end with undesirable consequences while reinforcing discrimination.

KEYWORDS: fairness, machine learning, sequential decision making, group representation

BIAS IN DECISIONS MADE BY MACHINES

Machine learning models developed using real-world data can inherit preexisting bias in the dataset. When applying the trained models to new instances, it may exhibit biases (e.g., with respect to sensitive attributes such as gender and race), either because the data collection process was biased, or because bias already exists in the underlying data. This bias is reflected in two ways in decision making: (1) it may lead to higher error variance for certain demographic groups and (2) it may lead to errors skewed in a particular direction for certain demographic groups, or both. For instance, COMPAS algorithm used by courts in the United States for recidivism prediction is biased against black defendants (Dressel & Farid, 2018); job searching platform XING ranks less qualified male applicants higher than female applicants who are more qualified (Lahoti, Weikum, & Gummadi, 2019); speech recognition products such as Amazon's Alexa and Google Home have accent bias against nonnative speakers (Harwell, 2018).

Moreover, decisions made about humans affect their actions. Bias in the decisions induces certain behavior, which is then captured in the dataset used to train future algorithms. This closed feedback loop becomes self-reinforcing and can lead to highly undesirable outcomes over time by allowing biases to perpetuate (O'Neil, 2016). Consider the speech recognition example given above where native speakers experience much higher quality than nonnative speakers. If this difference in experience leads more native speakers to use such products while driving away nonnative speakers, then over time the datasets used to train the speech recognition model may become even more skewed toward native speakers, with fewer and fewer nonnative samples. Without intervention, the resulting model will be more accurate for the former and less for the latter, which then reinforces their respective user experience.

To address the fairness issues highlighted above, one commonly used approach is imposing fairness constraints. Various notions of fairness have been proposed to formulate fairness mathematically (Chouldechova & Roth, 2018; Corbett-Davies & Goel, 2018.) and a majority of them require the (approximate) parity of certain statistical measure (e.g., positive classification rate, false positive rate, etc.) across different demographic groups.

LONG-TERM IMPACT OF FAIR MACHINE LEARNING

While the success of imposing fairness criteria in decision making has been shown in various domains (Hardt, Price, & Srebro, 2016), most of these studies are done using a static framework where only the immediate impact of the learning algorithm is assessed but not long-term consequences. In this section, we use two examples to highlight the long-term impact of fairness criteria when there is interaction between the decisions and the underlying population dynamics.

Long-Term Impact on Group Representation

Consider the speech recognition example given earlier, where machine learning models trained on data from multiple demographic groups inherit representation

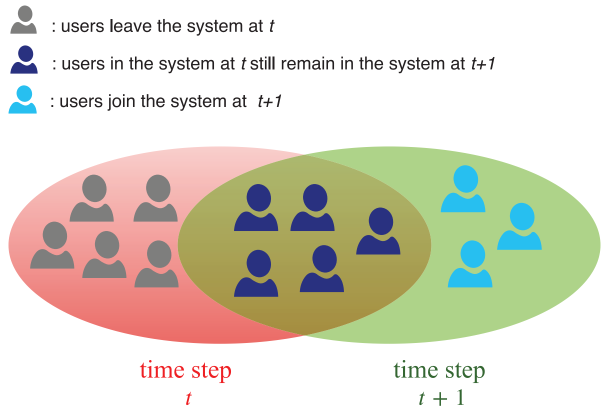

: users leave the system at t
: users in the system at t still remain in the system at t+1
: users join the system at t+1
time step
t
time step
t + 1

Figure 1. Retention dynamic model.

disparity in the data—the group contributing less to the training process suffers higher error. We now examine what happens when we impose a fairness criterion each time the model is used.

Suppose the user reaction is captured by a discrete-time retention/attrition dynamics (Zhang, Khalili, Tekin, & Liu, 2019) as follows: (1) users who experience low accuracy (or other forms of perceived mistreatment) have a higher probability of discontinuing their use of the models and (2) in each time step, certain new users will start using the algorithms (Figure 1). The number of users who choose to remain in the system is characterized by a retention rate. The algorithm tries to be fair by equalizing certain aspects of the model for different groups at each time step.

In Zhang et al. (2019), the long-term properties of decisions and group retention under a set of fairness criteria are characterized. It shows that as long as there is a mismatch between the fairness criterion and the factors driving user retention, the difference in (perceived) treatment will exacerbate group representation disparity in the long run. An inherent challenge is that this mismatch easily arises in various real-world scenarios because we typically have only incomplete and imperfect information on these factors.

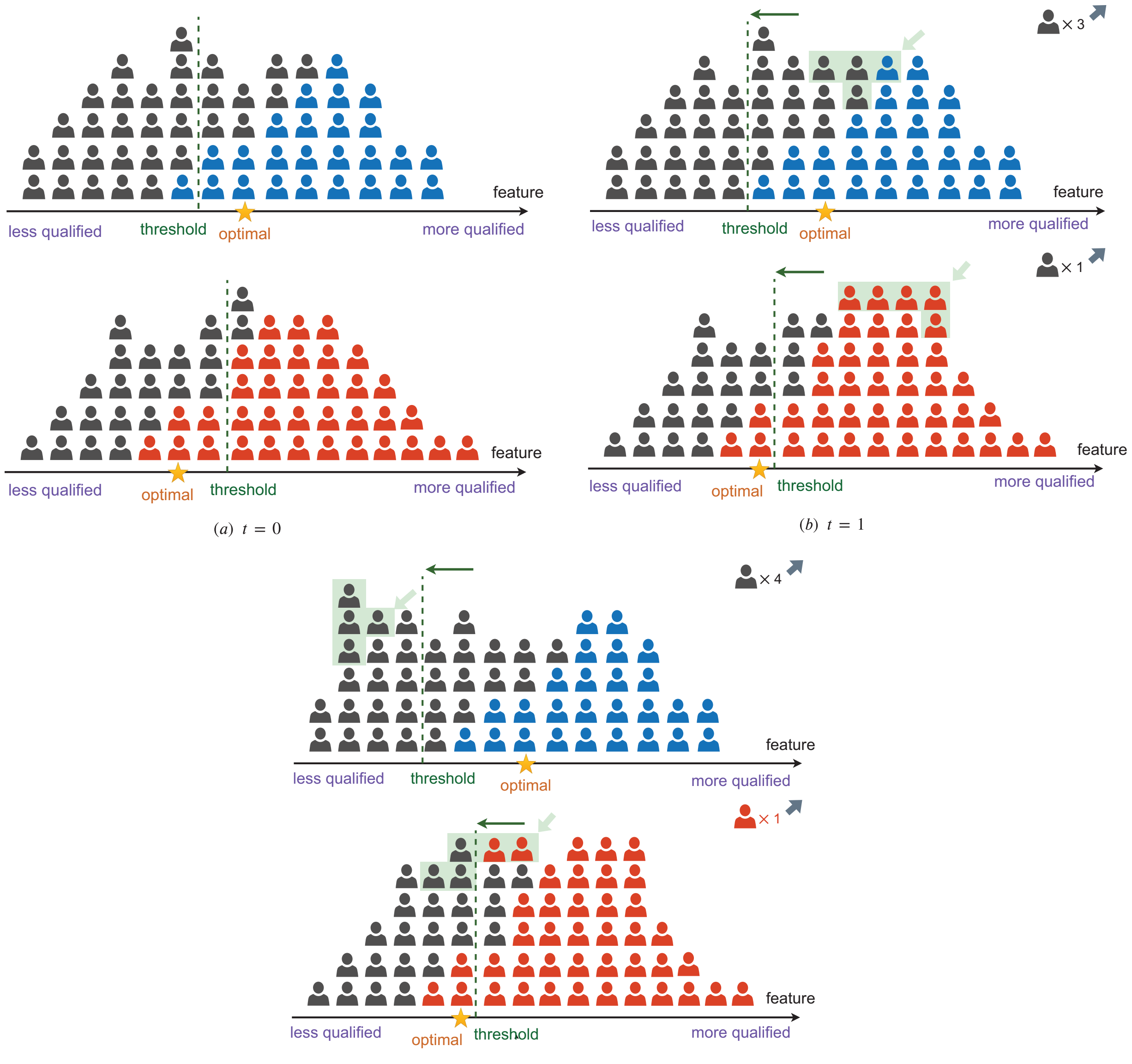
Consider the example of a dynamic model driven by model accuracy with two demographic groups (e.g., different racial groups with different accents). A user has a feature (e.g., extracted from his/her speech) that is observed by the decision maker and an underlying label 0 or 1 (e.g., not qualified/qualified for a certain reading job) is to be assigned. At each time, the decision maker finds a threshold for each group and makes classifications using the thresholds: assign “1” if the feature is above the threshold; assign “0” otherwise. As illustrated in Figure 2, where individuals from two groups are ordered by their features, individuals in gray (respectively, red/blue) are not qualified (respectively, qualified) and thus have label “0” (respectively, “1”). Two thresholds (green dashed line) are selected such that the same fraction of people is above it for both groups. Yellow star denotes the optimal threshold for each group without imposing fairness constraint, in minimizing the total classification error the group experiences; this error is calculated as the fraction of the population that is assigned incorrectly. At $t = 0$, the threshold is selected such that 57% of people are assigned label “1” for both groups. Since 9 (respectively, 7) out of 56 people from the blue (respectively, red) group are assigned incorrectly, the errors for blue and red groups are 16.07% and 12.5% respectively. At $t = 1$, because of the higher retention rate in the red group, the red group has 5 arrivals and 1 departure while the blue group has 4 arrivals and 3 departures, resulting in the red group having a higher proportion in the overall population. This drives the new threshold (assign 68% of people as label “1”) to move toward a direction in favor of the red group, that is, the threshold of red (respectively, blue) group is closer to (respectively, farther away from) its optimal threshold, further lowering its error (drop from 12.5% to $7/60 = 11.67\%$). In contrast, the blue group suffers a higher error (increase from 16.07% to $13/57 = 22.81\%$). This process continues monotonically as time goes on, leading to more (respectively, less) favorable decisions made for one (respectively, the other) group, diminishing the population of the disfavored group, and eventually causing it to disappear entirely from the system.

This example shows that group representation disparity may worsen if the aspects of the model we equalize differ from what actually affects user retention. Specifically, the imposed fairness criterion tries to equalize the fraction of people being above the threshold while it is accuracy that drives user retention. The guarantee of the former means difference in accuracy for the two groups, which eventually leads to the exacerbation of group representation disparity.

Long-Term Impact on Individuals

The previous example shows the potentially adverse long-term impact on group representation when imposing a fairness criterion that does not match user dynamics, while a user’s feature remains unaffected by either the error made by the algorithm or the user’s decision to stay or leave. We next show a second example, first studied in Liu, Dean, Rolf, Simchowitz, and Hardt (2018), where fairness criteria can lead to an adverse effect on the individual’s and entire populations’ features over time.

In this example, a lender decides whether or not to approve a loan application based on the applicant’s credit score (feature). To ensure fairness across different groups, the lender aims to achieve an identical loan approval rate (approving the same percentage of applications) in each group, or to achieve an identical true positive rate (approving the same percentage of applications among the qualified applicants). At the same time, for an applicant who has been issued a loan, his/her credit score improves if he/she repays and drops if he/she defaults (Figure 3). In Liu et al. (2018), it is shown that both equality criteria can potentially result in more loans issued to less qualified applicants in the group whose score distribution skews toward higher default risk. The lower repayment among these individuals causes their future credit scores to drop,



(a) $t = 0$

(b) $t = 1$

(c) $t = 2$

Figure 2. Illustration of the monotonic movement in decisions of two demographic groups.

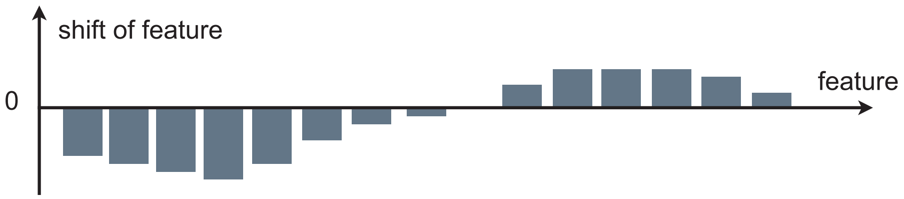


Figure 3. The average change of feature (credit score) on the applicant once it is assigned label "1" (issued the loan): when the loan is issued to individuals with higher (respectively, lower) credit score, because they are more likely to repay the loan (respectively, default), their credit scores can be improved (respectively, decreased) in average, that is, resulting in the positive (respectively, negative) shift of feature in *y*-axis.

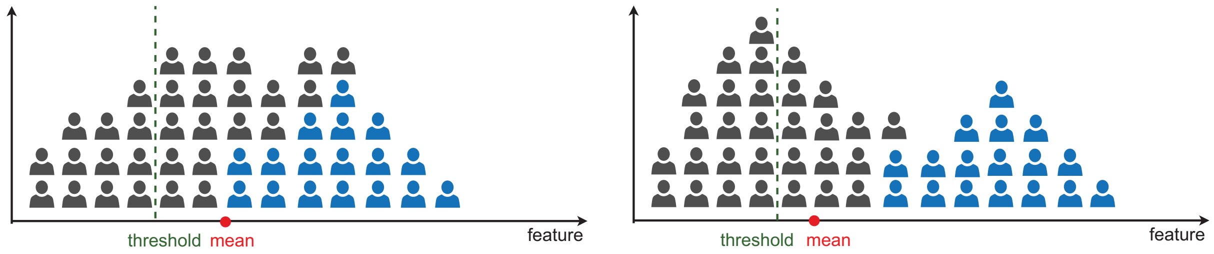


Figure 4. The average (red solid circle) feature (credit score) of the disadvantaged (less qualified) group decreases in the next time step (left, $t = 0$; right, $t = 1$): because the threshold is chosen such that unqualified applicants (gray) are over-issued the loan, based on Figure 3, credit score of these people will be decreased (negative shift of feature). Since credit score of a large portion of population decreases, the overall average credit score decreases.

which then causes the score distribution of that group to skew further toward high risk. In short, an attempt to improve immediate loan approval rate can inadvertently lead to worse opportunity in the long run for the very individuals the effort was designed to help (Figure 4).

Combination of Both Impacts

In practice, these two impacts can simultaneously exist, making a bad situation worse. Consider the lending example: once the lender starts to over issue loans to the less qualified group, the latter's score distribution will skew toward higher default risk. Over time, more people from this group may stop applying. The increased disproportionality between the two groups will then lead the lender to actually issue more loans (relatively) to the less qualified group if it continues to use the same type of fairness criterion. This will then lead its score distribution to skew more toward higher default risk over time. This is an example in which the two effects are both present and interact, resulting in compounding negative impact.

ROLE OF DYNAMIC MODELS IN IMPOSING FAIRNESS

The two examples show that imposing seemingly fair decision using an instantaneous criterion can lead to unintended consequences in the long run, for example, the extinction of one group in the system or deteriorating features of a population. They point to the fact that fairness has to be addressed with a good understanding of how users are affected by the algorithm and/or their perception of the algorithm, and how they may react to such perceptions. In other words, fairness cannot be defined in a one-shot problem setting without considering the long-run impact, and that long-run impact cannot be properly analyzed without understanding the underlying dynamics.

POTENTIAL MITIGATIONS AND CONCLUSIONS

Given a machine learning model, it is critical to ensure it can be accepted/trusted by users (Ribeiro, Singh, & Guestrin, 2016). In practice, decision makers should carefully inspect and measure how those affected perceive and react to the decisions made by such machine learning algorithms, any downstream effect following those perceptions and reactions, and do so over a sustained period of time. It is particularly important to capture and measure unintended consequences, that is, perceptions and reactions unanticipated by the algorithm designer, so as to inform adaptation and redesign of future algorithms.

In addition to the speech recognition example, for applications where the model at each time is trained on data from the current users in the system (either because the feature distribution of each group changes over time or because the decision maker has no access to historical user data), group representation should be carefully balanced to prevent models from being biased against a minority group. The fairness criterion should be chosen such that the factors driving user participation in the dynamic model can be equalized; for example, if user participation is driven by model accuracy, then decisions should be made such that different groups experience similar classification error. In reality, user reaction can be affected by a mixture of factors given different application contexts; thus, modeling user dynamics from real-world measurements and finding a proper fairness criterion based on the obtained model is an important research direction.

Similarly, to prevent the deterioration of features, decision making should carefully take into account future consequences and put in place measures intended to remedy any adverse effect. For instance, in the lending scenario, one cannot stop at simply issuing loans but must put in place aggressive repayment and other assistance programs for those receiving the loan to minimize default. In this sense, machine learning tools can be an enabling piece in a larger, more comprehensive system of policies and decision making, but should not be regarded as a self-sufficient solution on its own.

To conclude, good understanding of human factors should play a crucial role in promoting appropriate use of machine learning systems to produce fair, long-term outcomes rather than merely implementing fairness constraints. What has been discussed in this article are but a few examples of unintended consequences when human factors are not properly accounted for. Toward this end, we believe it is critical that more research be conducted to better understand not only existing biases in training datasets but also how users react to machine learning systems, including their tolerance of (perceived) unfairness, how it impacts their acceptance and rejection of such systems and how machine learning systems can be made more understandable so that decision makers can better anticipate outcomes.

REFERENCES

Chouldechova, A., & Roth, A. (2018). *The frontiers of fairness in machine learning*. Retrieved from https://arxiv.org/pdf/1810.08810.pdf

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning*. Retrieved from https://arxiv.org/pdf/1808.00023.pdf

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*, eaao5580.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, *29*, 3315-3323.

Harwell, D. (2018). *Amazon's Alexa and Google Home show accent bias, with Chinese and Spanish hardest to understand*. Retrieved from https://www.scmp.com/magazines/post-magazine/long-reads/article/2156455/amazons-alexa-and-google-home-show-accent-bias.

Lahoti, P., Weikum, G., & Gummadi, K. P. (2019). iFair: Learning individually fair data representations for algorithmic decision making. In *Proceedings of the 35th IEEE International Conference on Data Engineering, ICDE 2019*, *Macao, China, April 8-11, 2019* (pp. 1334-1345). doi:10.1109/ICDE.2019.00121

Liu, L., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018* (pp. 3156-3164). Retrieved from http://proceedings.mlr.press/v80/liu18c/liu18c.pdf

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Broadway Books.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). *New York, NY" ACM*.

Zhang, X., Khalili, M. M., Tekin, C., & Liu, M. (2019, December). *Group retention when using machine learning in sequential decision making: The interplay between user dynamics and fairness*. Paper to be presented at the 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, British Columbia, Canada.

Xueru Zhang *(xueru@umich.edu) received her BEng degree in electronic and information engineering from Beihang University (BUAA), Beijing, China, in 2015, and her MS degree in electrical and computer engineering from the University of Michigan, Ann Arbor, in 2016. She is currently pursuing her PhD degree in electrical and computer engineering at the University of Michigan. Her research interests include fairness and privacy in machine learning, distributed optimization, and sequential decision making. ORCID iD: https://orcid.org/0000-0002-0761-5943*

Mohammad Mahdi Khalili *(khalili@umich.edu) received his BS and MS degrees in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2013 and 2015, respectively, and his MS degree in applied mathematics from the University of Michigan, Ann Arbor in 2018. He is currently pursuing his PhD degree in electrical and computer engineering at the University of Michigan. His research interests include fairness in machine learning and the applications of mathematical economics in network security and privacy.*

Mingyan Liu *(mingyan@umich.edu, PhD in electrical engineering from the University of Maryland, College Park) is a professor and the Peter and Evelyn Fuss Chair of Electrical and Computer Engineering at the University of Michigan, Ann Arbor. Her interests are in sequential decision and learning theory, game theory and incentive mechanisms, with applications to large-scale networked systems. She is a fellow of the IEEE and a member of the ACM.*

DOI: 10.1177/1064804619884160
Article reuse guidelines: sagepub.com/journals-permissions