ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Dynamic fairness – Breaking vicious cycles in automatic decision making

Benjamin Paaßen Astrid Bunge Carolin Hainke Leon Sindelar
Matthias Vogelsang [*]

Bielefeld University - CITEC
Inspiration 1, 33619 Bielefeld - Germany

Abstract. In recent years, machine learning techniques have been increasingly applied in sensitive decision making processes, raising fairness concerns. Past research has shown that machine learning may reproduce and even exacerbate human bias due to biased training data or flawed model assumptions, and thus may lead to discriminatory actions. To counteract such biased models, researchers have proposed multiple mathematical definitions of fairness according to which classifiers can be optimized. However, it has also been shown that the outcomes generated by some fairness notions may be unsatisfactory.

In this contribution, we add to this research by considering decision making processes in time. We establish a theoretic model in which even perfectly accurate classifiers which adhere to almost all common fairness definitions lead to stable long-term inequalities due to vicious cycles. Only demographic parity, which enforces equal rates of positive decisions across groups, avoids these effects and establishes a virtuous cycle, which leads to perfectly accurate and fair classification in the long term.

Automatic decision-making via machine learning classifiers carries the promise of quicker, more accurate, and more objective decisions because automatic mechanisms do not foster animosity against any group [1, 2]. Yet, machine learning systems can indeed reproduce and exacerbate bias that is encoded in the training data or in flawed model assumptions [1, 2, 3, 4]. For example, the COMPAS tool, which estimates the risk of recidivism of defendants in the US law system prior to trial, has been found to have higher rates of false positives for Black people compared to white people and has thus been called unfair [5]. Similarly, a tool developed by Amazon to rate the résumés of job applicants assigned higher scores to men compared to women because successful applicants in the past had mostly been male [6]. Finally, multiple machine-learning-based credit scoring systems have emerged that reproduce historical biases and systematically assign lower credit scores to members of disenfranchised minorities [7].

In general, we consider scenarios where individuals $i \in \{1, \ldots, m\}$ in some population of size m apply for some positive outcome, such as a pre-trial bail, a job, or a loan, and a gatekeeper institution decides whether to grant that outcome, with the interest of accepting only those individuals who will "succeed" with that outcome, e.g. not commit a crime, succeed in their job for the company, or pay back a loan. To make that decision, the institution employs a binary classifier $f : \{1, \ldots, m\} \to \{0, 1\}$ that predicts whether to grant the outcome,

[*] Funding by the CITEC center of excellence (EXC 277) is gratefully acknowledged.

i.e. $f(i) = 1$, or not, i.e. $f(i) = 0$. Now, let $y_i \in \{0, 1\}$ denote whether an individual will succeed ($y_i = 1$) or not ($y_i = 0$). Then, the aim of the classifier is to maximize the share of the population where $f(i) = y_i$.

In our examples, we care how a certain *protected group* C is treated compared to everyone else. In general, we assume these protected groups to be pre-defined by society, e.g. via the EU charter of fundamental rights, which forbids discrimination based on sex, race, color, ethnic or social origin, religion, political opinion, and several other features [8]. Formally, let $C \subseteq \{1, \ldots, m\}$ be a protected group, let $m_c := |C|$, let $\neg C := \{1, \ldots, m\} \setminus C$, and let $m_{\neg c} := |\neg C|$. Then, the fairness notion corresponding to our last two examples is *demographic parity*, which requires that the rate of positive decisions is equal across groups, i.e. $\sum_{i \in C} \frac{f(i)}{m_c} = \sum_{i \in \neg C} \frac{f(i)}{m_{\neg c}}$ [4, 9].

Multiple authors have criticized demographic parity because it decreases accuracy if the base rate of successful people is different across groups [3, 9, 10]. Accordingly, Hardt et al. have proposed the notion of *equalized odds* which only requires an equal rate of positive decisions among the people who will succeed and the people who will not succeed [10], which corresponds to the fairness notion in the COMPAS example [5].

In addition distributive justice considerations, several authors have proposed notions of due process, in the sense that any classifier should be considered fair which performs decisions in a fair way [11]. In particular, several authors have argued that classifiers should not use features that code the protected group directly or indirectly [2, 4, 11, 12, 13]. Alternatively, Corbett and Goel have proposed a two-step classification process. First, a function $g : \{1, \ldots, m\} \to \mathbb{R}$ assigns a risk score to each individual, which should increase monotonously with the probability to be successful, i.e. $g(i) = \sigma(P(y_i = 1))$ for some monotonous function σ (a property also called *calibration* [14]). Second, the actual classifier only threshold the risk score, i.e. $f(i) = 1$ if $g(i) \geq \theta$ and $f(i) = 0$ otherwise for some fixed threshold $\theta \in \mathbb{R}$, thus holding everyone to the same standard [3].

In this contribution, we argue that even if a classifier is perfectly accurate and is fair according to all fairness notions except demographic parity, we may still obtain undesirable long-term outcomes. To do so, we establish a simple dynamical system which assumes that positive classifier decisions have positive impact on the future success rate of a group, which in turn leads to a higher chance for positive classifications and so on. We show that this positive feedback loop implies stable equilibria where a protected group receives no positive decisions anymore. We also show that imposing demographic parity breaks this feedback loop and introduces a single, stable equilibrium which exhibits perfect accuracy, equality, and fairness according to all notions.

Our model is inspired by prior work of O'Neil, who has investigated existing automatic decision making systems and found positive feedback loops which disadvantage protected groups [2]. However, O'Neil did not provide a theoretic model. Further, our work is related to prior research by Liu et al., who have analyzed one-step dynamics in a credit scoring scenario [14] but did not consider long-term outcomes. Third, Hu and Chen have previously analyzed a detailed economic model of the labor market, including long-term dynamics [15] and found that demographic parity leads to a desirable equilibrium. Finally, Mouzannar et al. generalized this work simultaneously and independently to us and analyzed a wide range of scenarios where acceptance decisions influence future qualifications [16]. Our work is similar to theirs, but we use a different model assuming continuous qualification variables, fixed institutional resources, and specific dynamics, which enables us to derive stronger conclusions.

1 Model

In our model, we assume that every individual i has an objective risk score q_i^t at time t which is drawn from an exponential distribution[1] with mean μ_c^t if $i \in C$ and with mean $\mu_{\neg c}^t$ otherwise. Further, we assume that the $n \leq m$ people with the highest score in each iteration are the ones which will be successful, i.e. $y_i = 1$ if and only if q_i^t is among the top n at time t. Accordingly, we obtain a perfectly accurate classifier if we use the scoring function $g^t(i) = q_i^t$ and set the decision threshold θ^t such that exactly the top n scores are above or equal to it. Note that our hypothetical classifier conforms to equalized odds because there are no misclassifications [10], fulfills the calibration, threshold, and accuracy requirements of Corbett and Goel [3], and does not need access to the group label, neither directly nor indirectly, thus conforming to all due process notions of fairness [2, 4, 11, 12, 13].

We estimate the overall number of people who receive a positive classification inside and outside the protected group via the expected values $\mathbb{E}[\sum_{i \in C} f(i)] = m_c \cdot \int_{\theta^t}^{\infty} \frac{1}{\mu_c^t} \cdot \exp(-\frac{q}{\mu_c^t}) dq = m_c \cdot \exp(-\frac{\theta^t}{\mu_c^t})$ and $\mathbb{E}[\sum_{i \in \neg C} f(i)] = m_{\neg c} \cdot \exp(-\frac{\theta^t}{\mu_{\neg c}^t})$[2].

We finally assume that the mean for a group improves with a higher rate of positive classifier decisions in the previous time step according to the following equation.

$$\begin{pmatrix} \mu_c^{t+1} \\ \mu_{\neg c}^{t+1} \end{pmatrix} = (1 - \alpha) \cdot \begin{pmatrix} \mu_c^t \\ \mu_{\neg c}^t \end{pmatrix} + \beta \cdot \begin{pmatrix} \exp(-\frac{\theta^t}{\mu_c^t}) \\ \exp(-\frac{\theta^t}{\mu_{\neg c}^t}) \end{pmatrix} \tag{1}$$

where the decision threshold θ^t is selected as the numeric solution to the equation $n = m_c \cdot \exp(-\frac{\theta^t}{\mu_c^t}) + m_{\neg c} \cdot \exp(-\frac{\theta^t}{\mu_{\neg c}^t})$, where the parameter $\alpha \in [0, 1]$ quantifies the score fruction an individual loses in each time step ("leak reate"), and where the parameter $\beta \in \mathbb{R}^+$ quantifies the score an individual gains for for a positive classifier decision. Figure 1 (left) visualizes the dynamical system.

Note the connections of our model to the real-world examples mentioned before. In credit scoring, q_i^t would correspond to the credit score, i.e. the ca-

[1] Note that our qualitative results can be generalized to other distributions, such as Gaussian or Pareto. We select the exponential distribution here because it only has a single parameter and thus is easier to analyze. You can find the full analysis in the appendix at `https://arxiv.org/abs/1902.00375`.

[2] We consider each classifier decision as a Bernoulli trial with success probability $P = \int_{\theta^t}^{\infty} \frac{1}{\mu_c^t} \cdot \exp(-\frac{q}{\mu_c^t}) dq$, yielding a binomially distributed random variable $\sum_{i \in C} f(i)$ with expected value $m_c \cdot P$ and variance $m_c \cdot P \cdot (1 - P)$. Note that the variance gets close to zero if P is small itself, such that the expected value is a precise estimate for sufficiently small n.

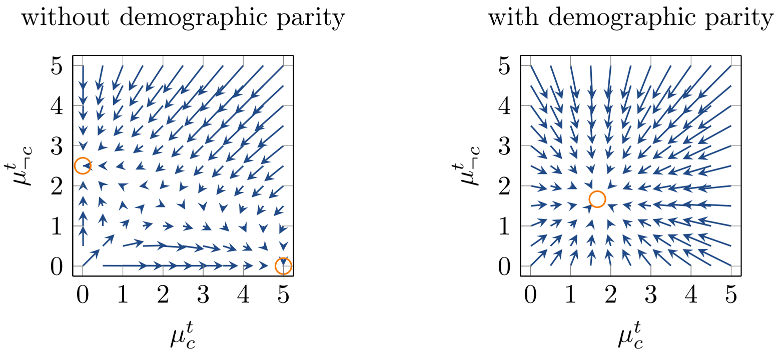

Figure 1: An illustration of the dynamical system model from Equation 1 for a population with $m_c = 100$, $m_{\neg c} = 200$, $n = 50$ successful people, leak rate $\alpha = 0.5$, and score $\beta = 5$. Equilibria are highlighted with circles. Left: The model without demographic parity requirement, exhibiting undesirable stable equilibria at the coordinate axes. Right: The model with demographic parity, exhibiting a single stable equilibrium on the diagonal.

pability of an individual to pay back a loan. We would plausibly assume that the score increases with positive classifier decisions because individuals who get a loan have additional financial resources at their disposal and can use those to add wealth to their group [14]. Further, we would assume a nonzero leak rate α because individuals need to cover their expenses which may negatively affect their capability to pay back a loan.

If we apply our model to pre-trial bail assessment, the score q_i^t would assess the likelihood of a defendant to not commit a crime until trial. Here, we would assume that the score decreases with negative classifier decisions because incarcerating people from a community may cut social ties and deteriorate trust in the state, leading to a higher crime rate [2]. This effect can be modeled by a nonzero leak rate α and a positive score β.

Also note that our model is not necessarily realistic but shows that there exist contexts where even perfect classifiers can exhibit stable long-term inequality. We show that context can matter, not that every context conforms to our model.

If we analyze the equilibria of this system, we first note that $\lim_{\mu_c^t \to 0} \mu_c^{t+1} = \lim_{\mu_c^t \to 0} (1-\alpha) \cdot \mu_c^t + \beta \cdot \exp(-\frac{\theta^t}{\mu_c^t}) = 0$, i.e. $\mu_c^* = 0$ is a fix point. Further, $\lim_{\mu_c^* \to 0} \exp(-\frac{\theta^*}{\mu_c^*}) = 0$, i.e. no person from the protected group is above the threshold at that fix point. Accordingly, we can compute the fix point threshold θ^* only for the non-protected group, i.e. $\theta^* = \mu_{\neg c}^* \cdot \log(\frac{m_{\neg c}}{n})$. By plugging this into the fix point equation $\mu_{\neg c}^* = (1-\alpha) \cdot \mu_{\neg c}^* + \beta \cdot \exp(-\frac{\theta^*}{\mu_{\neg c}^*})$ we obtain $\mu_{\neg c}^* = \frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}}$, which yields $\mu_{\neg c}^* = 2.5$ for our example in Figure 1 (left). At this fix point, we obtain a Jacobian of Equation 1 which is $1 - \alpha$ times the identity matrix, i.e. both eigenvalues have an absolute value < 1 for $\alpha > 0$, implying stability. In Figure 1 (left) we also see that the basin of attraction is the entire region above the diagonal, i.e. whenever we start with slight inequality in favor of the non-protected group, this inequality will get amplified.

In summary, we have shown that, for our exponential distribution model, there are always undesirable and stable equilibria in which μ_c^t degenerates to zero and the non-protected group receives all positive outcomes. This begs the question: Can we break this undesirable dynamic? Indeed, we can, using demographic parity.

2 Demographic Parity Dynamics

Demographic parity requires equal acceptance rates across groups, i.e. $\exp(-\frac{\theta_c^t}{\mu_c^t}) = \exp(-\frac{\theta_{\neg c}^t}{\mu_{\neg c}^t}) = P$ for some acceptance rate P and group-specific thresholds θ_c^t and $\theta_{\neg c}^t$. We obtain P as solution of the threshold equation $n = m_c \cdot P + m_{\neg c} \cdot P$, i.e. $P = \frac{n}{m_c + m_{\neg c}} = \frac{n}{m}$. By plugging this result into our fix point equation we obtain $\mu^* = \mu_c^* = \mu_{\neg c}^* = (1 - \alpha) \cdot \mu^* + \beta \cdot P = \frac{\beta}{\alpha} \cdot \frac{n}{m}$, which yields $\mu^* = 5/3$ for our example in Figure 1 (right). For this fix point we obtain a Jacobian of Equation 1 of $1 - \alpha$ times the identity matrix, implying stability.

Overall, demographic parity ensures that the mean for every group converges to the same point, such that the thresholds θ_c^t and $\theta_{\neg c}^t$ become equal as well. This, in turn, implies that selecting the top-scored people in each group corresponds to selecting the top-scored people in the entire population, implying a classifier that is perfectly accurate *and* conforms to all notions of fairness, including demographic parity.

3 Conclusion

In this contribution, we have analyzed a simple dynamic model for automatic decision making. In particular, our model assumes that people should receive a positive classifier decision only if their objective risk score is in the top, that the means of the score distribution differ between the protected group and everyone else, and that positive decisions improve the mean for the group in the next time step. This feedback loop becomes a vicious cycle in which even a perfectly accurate classifier conforming to almost all fairness notions leads to stable inequality. Fortunately, we can break this vicious cycle by imposing democratic parity which instead leads to an equilibrium with perfectly accurate, equal, and fair classification.

At present, our analysis is limited to a theoretical model assuming an exponential distribution and a simple dynamic model. However, we note that generalizations to other distributions are possible. Further, we note that our findings are consistent with practical application scenarios [2] and other theoretic studies [15, 16].

Overall, we conclude that our findings give reason to re-think notions of fairness in terms of mid- and long-term outcomes and reconsider demographic parity as a helpful intervention whenever decision making systems are embedded in vicious cycles. Otherwise, even well-intended and well-constructed systems may stabilize and exacerbate inequality.

References

[1] Cecilia Munoz, Megan Smith, and DJ Patil. Big data: A report on algorithmic systems, opportunity, and civil rights, 2016.

[2] Cathy O'Neil. *Weapons of Math Destruction: How big data increases inequality and threatens democracy.* Crown, Random House, New York City, NY, USA, 2016.

[3] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. In *Tutorial at ICML 2018*, 2018.

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS 2012*, pages 214–226, 2012.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *Pro Publica*, 2016.

[6] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018.

[7] Rachel O'Dwyer. Algorithms are making the same mistakes assessing credit scores that humans did a century ago. *Quartz*, 2018.

[8] European Union Agency for Fundamental Rights. *Article 12, Non-Discrimination*, volume C 326/391. 2012.

[9] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, 2017.

[10] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS 2016*, pages 3315–3323, 2016.

[11] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS 2016 Workshop "ML and the Law"*, 2016.

[12] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *NIPS 2017*, pages 656–666. 2017.

[13] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS 2017*, pages 4066–4076, 2017.

[14] Lydia Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause, editors, *ICML 2018*, pages 3156–3164, 2018.

[15] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *WWW 2018*, pages 1389–1398, 2018.

[16] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *FAT* 2019*, pages 359–368, 2019.