Context-conscious fairness in using machine learning to make decisions

Michelle Seng Ah Lee (University of Oxford; michelle.lee@oii.ox.ac.uk)

DOI: 10.1145/3340470.3340477

Abstract

The increasing adoption of machine learning to inform decisions in employment, pricing, and criminal justice has raised concerns that algorithms may perpetuate historical and societal discrimination. Academics have responded by introducing numerous definitions of "fairness" with corresponding mathematical formalisations, proposed as one-size-fits-all, universal conditions. This paper will explore three of the definitions and demonstrate their embedded ethical values and contextual limitations, using credit risk evaluation as an example use case. I will propose a new approach - context-conscious fairness - that takes into account two main trade-offs: between aggregate benefit and inequity and between accuracy and interpretability. Fairness is not a notion with absolute and binary measurement; the target outcomes and their trade-offs must be specified with respect to the relevant domain context.

Introduction

Machine learning (ML) is increasingly used to make important decisions, from hiring to sentencing, from insurance pricing to providing loans. There has been an explosion of publications of new guidelines and principles of Artificial Intelligence (AI) from governments (Holden & Smith, 2016), international organisations (European Commission AI HLEG, 2019), professional associations (IEEE Global Initiative, 2018), and academic institutions (Grosz, 2015). Each of these documents references the imperative for AI to treat people fairly by avoiding discrimination based on legally protected characteristics, such as race, gender, and sexual orientation. However, none of them specifies a framework or methodology to detect and correct unfair outcomes, especially given history of discrimination in our systems and societies that predates the introduction of algorithmic decision-making. Given a bias, people-based processes may arrive at different decisions. AI, by contrast, can replicate an identical bias at-scale, crystallising the bias and removing the outcome ambiguity associated with human decision-making. This is especially concerning in domain areas with documented historical discrimination, as AI can exacerbate any underlying societal problems and inequalities. For example, discriminatory lending has been a contentious problem. In the United States, 'redlining,' risk inflation of minority-occupied neighbourhoods, impeded African Americans from obtaining mortgage (Nelson, Winling, Marciano, & Connolly, 2016). Analysis of 2001-2009 UK consumer credit data on 58,642 households found that non-white households are less likely to have financing (Deku, Kara, & Molyneux, 2016). To counteract this, new techniques have been introduced for *pre-processing* (purging the data of bias prior to training the algorithm) and for *post-processing* (bias correction in predictions after algorithm build). In both cases, solutions are based on a one-size-fits-all condition of *fairness*, regardless of the context.

There is an overall consensus that AI systems and technologies should be required to make fair decisions, yet the binary categorisation of an algorithm as either fair or unfair belies the underlying complexities of each use case. In the next section, I will use discriminatory lending as an example to demonstrate the shortcomings in existing attempts at formalising fairness in machine learning predictions. Different issues may rise in other domains, such as employment, pricing, or criminal justice; however, the focus on one use case will help bring to light the real-life considerations of each fairness definition. In Section 3, I will propose a new approach focusing on quantifying the contextual trade-offs of each algorithm so that the decision-makers can select a model that best captures the risk profile of each specific case.

Ethics of fairness definitions and their limitations

If the outcome disparity in loan decisions is simply a factor of significant features (e.g., difference in income), arguably the variance in outcome may be consistent with the underlying distributions. Given that minority borrowers are indeed riskier, from an expected consequentialist perspective, overall societal welfare would increase with a more accurate prediction of default risk. If we subscribe to this, the machine learning model is fair insofar as it most accurately predicts risk and gives each individual the loan and the interest rate that he or she deserves, by foreseeing the consequences of a loan approval. However, the outcome disparity could be a reflection of inequality in other markets (e.g. in labour markets), which makes minority incomes more volatile. Given the evidence in past literature of historical discrimination based on race and gender in mortgage lending decisions in the US (Kendig, 1973), a machine learning algorithm can self-perpetuate a reproduction of past inequalities, inaccurately inflating the risk of minority borrowers. The extent of this inaccuracy is difficult to determine; it is impossible to know whether those who were denied a loan would have defaulted. The scarcity and potential bias of historical data on previously financially marginalised groups hinders the statistical modelling of their credit-worthiness. If historical evaluations of minority credit risk are inaccurate, a machine learning model trained on a biased data set necessarily produces a sub-optimal result. I will discuss three of the main approaches that seek to define whether or not an outcome is unfair.

Demographic parity

A strictly egalitarian approach mandates *equal outcome* for each racial group. A related statistical definition is *demographic parity*, a population-level metric that requires the outcome to be independent of the protected attribute. Formally, with Ŷ as the predicted outcome and A as a binary protected attribute, we have:

$$P(\hat{Y}|A=0) = P(\hat{Y}|A=1) \qquad (1)$$

While this metric would ensure the equally proportional outcome independent of race, it is ineffective where disproportionality in outcomes can be justified by non-protected, non-proxy attributes, as this can lead to reverse discrimination and inaccurate predictions (Gajane, 2017). Fuster et al. have shown that there is a difference in income distributions between racial groups (Fuster, Goldsmith-Pinkham, Ramadorai, & Walther, 2017). As income has a reasonably inferential relationship to credit risk, it cannot be considered as a simple proxy attribute of protected characteristics. In short, implementing this measure would unfairly give credit to minorities with low income.

Equalised opportunity

According to a Rawlsian approach to distributive justice, regardless of whether policymakers have the moral imperative to correct past wrongs of systematic discrimination, it is important to prioritise the protection of the most vulnerable of the population. The Max-Min social welfare function maximises the welfare of those who are worst-off (Rawls, 1971). In contrast to equal outcome, Rawls' Difference Principle asserts the need for *equality in opportunity*.

Hardt, Sbrero and Price proposed a statistical metric of "equalised opportunity" that focuses on the true positives: given a positive outcome, the prediction is independent of the protected attribute. Formally, again with Ŷ as the predicted outcome, A as a binary protected attribute, and Y as the actual outcome, we have:

$$\begin{aligned} &P(\hat{Y}=1|A=0,Y=1) \qquad (2)\\ &= P(\hat{Y}=1|A=1,Y=1) \end{aligned}$$

Unfortunately, equalised opportunity metric is also problematic because it fails to address discrimination that may already be embedded in the data (Gajane, 2017). Gender and race are outside of one's control, and following the logic of Dworkin's theory of Resource Egalitarianism, no one should end up worse off due to bad luck, but rather, people should be given differentiated economic benefits as a result of their own choices (Dworkin, 1981). The credit risk market does not exist in a vacuum; while people can affect their credit scores and income to a certain extent, e.g. by building their credit history or improving employable skills, it is impossible to isolate similar variables from the impact of their upbringing, discrimination in other markets, and historical inequalities entrenched in the data.

Counterfactual fairness

A challenge to the previous two fairness metrics is that they do not take seriously the notion of causality. Statistical relationships are derived from correlations rather than an inferential and directed model. David Lewis (1973) introduced the counterfactual approach of theorising on the cause and effect (Lewis, 1973). *Counterfactual fairness*, one of the most recently introduced definitions, attempts to apply this theory to assert that a decision is fair if it is the same in the actual world as it would be in a counterfactual world where the individual belonged to a different demographic group (Kusner, Loftus, Russell, & Silva, 2017). Given a causal model with latent background variables U, non-proxy non-protected features X, and the protected attribute A, Ŷ as the predicted outcome, and Y as the actual outcome, we have:

$$\begin{aligned} &P(\hat{Y}_{A\leftarrow a}(U) = Y | X = x, A = a) \qquad (3) \\ &= P(\hat{Y}_{A\leftarrow a'}(U) = y | X = x, A = a) \end{aligned}$$

However, this approach requires a formalisation of causal models, which are useful abstractions but often infeasible without strong assumptions, simplifications, and robust constraints. In reality, relationships between variables are intertwined, complex, or not fully understood. For example, how would we isolate the impact of one's race on credit-worthiness from the impact of one's education, job, neighbourhood, and other features of one's upbringing? There may also be challenges of reverse causality: e.g., race may impact one of the non-protected predictors, e.g. if employers are discriminatory, so that an applicant's low income ends up hindering both the probability of loan approval and his or her ability to repay a loan.

Implementing a "fair" credit risk algorithm is difficult not just because there are competing theoretical approaches to fairness, but because they all assume that there is a one-size-fits-all definition of fairness. What seems to be required is a new approach that builds on the previous ones but makes the most of the context-dependency of the data available and hence of the relational nature of fairness (*a*) is fair - not absolutely - but in relation to (*b*). I introduce a possible proposal in the next section.

Proposed solutions

While the implications of unequal outcome in mortgage lending are troubling, the revelation is also an opportunity for policymakers to redefine anti-discrimination policies. Consider the alternatives to a machine learning model. Human-driven decision-making is mired in cognitive biases, as shown by a field experiment on discrimination in the labour market, in which white-sounding names received 50% more callback than black-sounding names (Bertrand & Mullainathan, 2003). Traditional credit risk models, e.g. a scorecard, are also not exempt from selection and counterfactual biases of a machine-learning model. The accusation of discrimination in mortgage lending long predates the introduction of machine learning; bias can be embedded in a process (Kendig, 1973).

In reality, the alternative to a machine learning model may be a worse model. Therefore, I propose a benchmarking exercise to map the change in the two trade-offs, while increasing the model complexity. While no model, pre-processed or post-processed, may eliminate all the discriminatory impact on minority groups, quantifying the benefits and risks would give actionable insights to the decision-maker on which model best reflects his or her values and risk profile. Algorithmic processing of credit worthiness would not translate into automatic decision making but would empower a better evaluation of each case under scrutiny.

The shift of focus from discriminatory intent to discriminatory impact in scholarship is mirrored in legal rulings. "Fairness through Unawareness" approach - not using the protected characteristic - is invalidated with ML models that can triangulate protected information.[1] Even if a lender is not explicitly considering race in calculating credit risk, a machine learning model may nonetheless incorporate

[1] See: (Dwork, Hardt, Pitassi, Reingold, & Zemel, 2011) for an extensive critique.

racial information in its prediction. Given legal decisions in the UK (Lowenthal, 2017) and in the US (Baum & Stute, 2015), discriminatory impact is illegal regardless of intent. Both courts emphasise a contextual exception: disparate impact is allowed if it is crucial to a legitimate business requirement. One of such requirements may be the need to predict accurately default to allow for greater financial inclusion, which - given discriminatory history - may be at the expense of some minority groups.

Trade-off 1: Aggregate benefit vs. individual inequity

Credit market welfare can be measured by the consumer surplus generated. Automated underwriting improves the overall accuracy of default predictions (Hacker & Wiedemann, 2017). This increases the market information available to the lenders and lowers their overall risk, empowering them to give credit to those who were previously unable to access it. The resulting increase in financial inclusion adds to the overall consumer surplus.

On an individual level, however, there are winners and losers: the borrowers who would not have received a loan under traditional technology who now successfully secure one; and the potential borrowers who would have received a loan before but are deemed too high risk under the new technology. A troubling finding of Fuster et al. is that the losers are disproportionately from minority racial groups, likely as a reflection of historical discrimination that is inaccurately inflating the risk of minority borrowers (Fuster et al., 2017). Meanwhile, the various attempts at an active correction of this “unfairness”, synthesised in Section 2, have only obfuscated the criteria for fairness, as they do not provide a satisfactory framework to determine which definition is most appropriate for a given use case. Mapping of this trade-off between aggregate benefit (e.g. increased financial inclusion) and inequity (e.g. negative impact on minority borrowers) would allow the lender to identify the most appropriate algorithm to model default risk.

Figure 1 is an indicative visualisation of this trade-off. There are two assumptions that should be tested: 1) aggregate benefit improves with the complexity of the model and

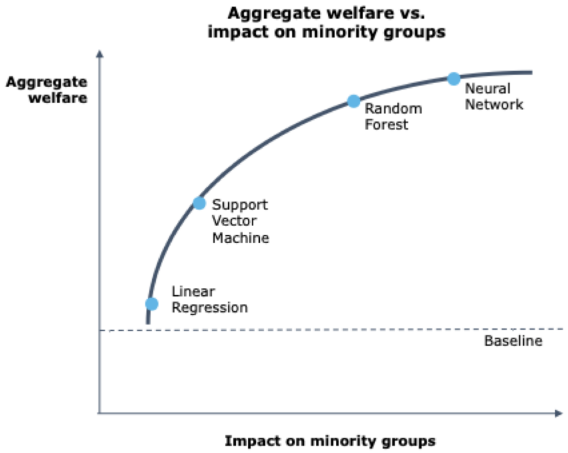

Figure 1: Benchmarking models for trade-off in aggregate group benefit vs. the scale of negative impact on protected groups

2) The increase in benefit is at the expense of the welfare of the protected group. The first assumption is reasonable for most complex models in which an implementation of a machine learning algorithm is being considered. If the true relationship is linear, then the out-of-sample accuracy would decrease with additional complexity, but this will be rare in many real-life modelling scenarios. The second assumption would depend on the joint distribution of the protected characteristic and the outcome being predicted. In many cases, the risk of discrimination exists because of the statistical disparity in outcomes between groups.

Rather than enforcing a single definition of fairness, this allows a context-conscious analysis of which model best serves the customers or society. This exercise should give a decision-maker an insight into the model and the trade-off with which he or she is the most comfortable. Once these trade-offs are recognised and quantified, it is important to consider whether the inter-group outcome disparity is fair and justified.

Trade-off 2: Accuracy vs. interpretability

Historically, lenders have focused on data that theoretically relate to outcome; for example, the debt-to-income ratio and past payments are indicators of default risk. With Big Data analytics, firms are beginning to incorporate non-traditional data types into their algorithms as proxies of risk. An extreme case is the use

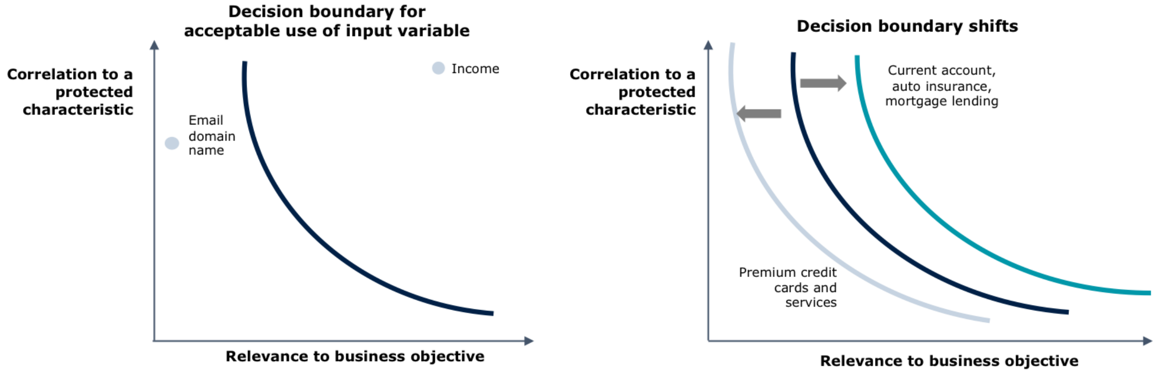
Decision boundary for acceptable use of input variable
Correlation to a protected characteristic
Income
Email domain name
Relevance to business objective
Decision boundary shifts
Correlation to a protected characteristic
Current account, auto insurance, mortgage lending
Premium credit cards and services
Relevance to business objective

Figure 2: Decision boundary for acceptable use of feature

of Chinese citizens' Internet browsing history, location, and payment data to calculate credit risk (Koren, 2016). While the justification is to open up access to credit to those without credit history, this data use risks unfairly marginalising people based on their background, in uninterpretable and opaque ways. Previously, a prospective borrower could build his or her credit history by paying bills on time and expect a positive impact on the probability of loan approval. With Big Data and machine learning, this expectation may be unfounded.

A decision-maker must justify whether the outcome disparity is due to legitimate and non-discriminatory difference in the outcome distributions. Figure 2 visualizes a possible decision boundary for whether or not an input variable should be used in a model, based on its role as a potential proxy for a protected characteristic. Given Supreme Court decisions in the UK (Lowenthal, 2017) and the US (Baum & Stute, 2015), even if a variable is correlated to a protected feature, there may be reasonable grounds to use it if the differences are crucial to a legitimate business requirement. This decision boundary may shift depending on the context. The drivers of decision-making in providing essential products, such as current account, car insurance, or mortgage, may be subject to higher scrutiny than the rationale for offering premium credit cards. This ensures that the decision to include features correlated to protected characteristic is carefully considered within the context of the regulated domain and the potential impact on the customers.

Policymakers should consider this trade-off between accuracy and interpretability to limit what data can be used and what models can be built. For provision of essential products, e.g. current accounts, criminal justice decisions, and car insurance, that have a significant impact on people's lives, policymakers may need to ensure that all features included in the variable have a strong inferential relationship with the outcome rather than simply for predictive correlation.

Conclusion

The introduction of ML into lending models presents a new opportunity for greater accuracy in predicting default and, therefore, greater financial inclusion. However, the prevalence of ML models in crucial decision-making processes has brought to light the intricate ways in which discrimination can be perpetuated through these technologies. Fairness is a complex feature of the world, and no single definition can mitigate the risk of unfair treatment.

At the same time, ML models are auditable for fairness. The opportunity lies in the computational and systematic decision-making of ML. A more context-conscious approach of benchmarking the trade-offs between aggregate benefit and individual inequity and between accuracy and interpretability would force the decision-maker to identify the model that is most suitable to each case. The risk of discrimination in ML is undeniable; however, it presents an opportunity to consider more closely what we value as a society and to pursue fair treatments and decisions that can be enforced in our markets.

References

Baum, J. S. J. J., D., & Stute, D. (2015). Supreme court affirms fha disparate impact claims.

Bertrand, M., & Mullainathan, S. (2003, July). *Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination* (Working Paper No. 9873). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w9873 doi: 10.3386/w9873

Deku, S. Y., Kara, A., & Molyneux, P. (2016). Access to consumer credit in the uk. *The European Journal of Finance*, *22*(10), 941-964. Retrieved from https://doi.org/10.1080/1351847X.2015.1019641 doi: 10.1080/1351847X.2015.1019641

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2011). Fairness through awareness. *CoRR*, *abs/1104.3913*. Retrieved from http://arxiv.org/abs/1104.3913

Dworkin, R. (1981). What is equality? part 1: Equality of welfare. *Philosophy and Public Affairs*, *10*(3), 185–246.

European Commission AI HLEG. (2019). *Draft ai ethics guidelines for trustworthy ai* (Tech. Rep.). European Commission High-Level Expert Group on Artificial Intelligence (AI HLEG). Retrieved from https://ec.europa.eu/futurium/en/node/6044

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2017, 01). Predictably unequal? the effects of machine learning on credit markets. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3072038

Gajane, P. (2017). On formalizing fairness in prediction with machine learning. *CoRR*, *abs/1710.03184*. Retrieved from http://arxiv.org/abs/1710.03184

Grosz, R. A. E. H. A. M. T. M. D. M. Y. S., Barbara. (2015). *Artificial intelligence and life in 2030* (Tech. Rep.). Stanford University. Retrieved from https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf

Hacker, P., & Wiedemann, E. (2017). A continuous framework for fairness. *CoRR*, *abs/1712.07924*. Retrieved from http://arxiv.org/abs/1712.07924

Holden, J., & Smith, M. (2016). *Preparing for the future of ai* (Tech. Rep.). U.S. Executive Office of the President, National Science and Technology Council Committee on Technology. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

IEEE Global Initiative. (2018). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems.* Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Kendig, D. (1973). Discrimination against women in home mortgage financing. *Yale Review of Law and Social Action*, *3*(2).

Koren, J. R. (2016, Jul). What does that web search say about your credit? Retrieved from https://www.latimes.com/business/la-fi-zestfinance-baidu-20160715-snap-story.html

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017, March). Counterfactual Fairness. *arXiv e-prints*, arXiv:1703.06856.

Lewis, D. (1973). Causation. *Journal of Philosophy*, *70*(17), 556–567.

Lowenthal, T. (2017). Essop v home office: Proving indirect discrimination.

Nelson, R. K., Winling, L., Marciano, R., & Connolly, N. (2016). Mapping inequality. *American Panorama*. Retrieved from https://dsl.richmond.edu/panorama/redlining/#loc=5/36.721/-96.943&opacity=0.8&text=intro

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Michelle Seng Ah Lee
michelle.lee@oii.ox.ac.uk

Michelle is an MSc candidate in Social Data Science at the Oxford Internet Institute and a Research Assistant at the Digital Ethics Lab, supervised by Professor Luciano Floridi. Her research focuses on fairness in machine learning algorithms and their trade-offs on aggregate and individual levels. She is also interested in the regulatory and ethical considerations of bias and discrimination in artificial intelligence (AI).