A Performance-preserving Fairness Intervention for Adaptive Microfinance Recommendation

Robin Burke
University of Colorado, Boulder
Boulder, Colorado, USA
robin.burke@colorado.edu

Pradeep Ragothaman
Kiva Microfunds
San Francisco, California, USA
pradeepr@kiva.org

Nicholas Mattei
nsmattei@tulane.edu
Tulane University
New Orleans, Louisiana, USA

Brian Kimmig
Reddit, Inc.
San Francisco, California, USA
brian.kimmig@gmail.com

Amy Voida
amy.voida@colorado.edu
University of Colorado, Boulder
Boulder, Colorado, USA

Nasim Sonboli*
University of Colorado, Boulder
Boulder, Colorado, USA
nasim.sonboli@colorado.edu

Anushka Kathait
University of Colorado, Boulder
Boulder, Colorado, USA
anushka.kathait@colorado.edu

Melissa Fabros
melissa.fabros@gmail.com
Independent Scholar
San Francisco, California, USA

ABSTRACT

Recommender systems are among the most widely-deployed and frequently-encountered machine learning systems for the general public. The fairness properties of such systems have been investigated by scholars and industry practitioners alike. However, despite a growing literature in this area, there are very few reports that both discuss deployed fairness-promoting interventions in detail and describe specific findings of those deployments. We describe a project conducted at Kiva Microfunds, a non-profit organization that seeks to promote global financial inclusion. In service of this mission, we sought increase the exposure of underfunded loans. We describe the adaptive recommendation framework used on Kiva's website, the intervention we performed, the A/B testing used to evaluate its impact, and our findings. Specifically, we show that contrary to the assumptions of much prior research, the inclusion of items specifically for the purposes of promoting fairness objectives has no statistically-significant impact on key performance indicators, indicating that there may be greater latitude for the inclusion of fairness objectives in recommender systems than previously thought.

KEYWORDS

fairness, recommender systems, multi-arm bandit, machine learning, financial inclusion

*Now at Tufts University

OARS '22, August, Washington, DC

ACM ISBN unknown…$xx.00
https://doi.org/xx.xxxx/yyyyyyy.yyyyyy

ACM Reference Format:
Robin Burke, Pradeep Ragothaman, Nicholas Mattei, Brian Kimmig, Amy Voida, Nasim Sonboli, Anushka Kathait, and Melissa Fabros. 2022. A Performance-preserving Fairness Intervention for Adaptive Microfinance Recommendation. In *OARS 22: 2nd Workshop on Online and Adaptive Recommender Systems.* ACM, New York, NY, USA, 6 pages. https://doi.org/xx.xxxx/yyyyyyy.yyyyyy

1 INTRODUCTION

Recommender systems are widely deployed in media streaming, social networking, e-commerce and a host of other applications [3]. Millions of users encounter these automated personalized decision support systems every day. Because of their broad impact, the fairness properties of recommender systems have come under scrutiny, just as has been the case for machine learning systems more generally [5].

While there is a substantial literature in the area of fairness-aware recommender systems (see the survey in [7]), there is very little published research that reports on actual deployment of fairness-aware recommender systems to users, i.e., systems that have been re-designed with an intervention to promote some definition of fairness in addition to personalization. There are some exceptions, for example [12], but in general, organizations are reluctant to go public with fairness problems that they may have identified in their products.

This gap in the literature is unfortunate, as it means that this surge of interest in the fairness properties of recommender systems is in danger of becoming disconnected from the real-world problems it is meant to solve. As Cramer, Vaughn and colleagues noted in their 2019 tutorial [6], the research community has tended to adopt an idealized view of fairness, e.g., positive parity between protected and unprotected groups, independent simultaneous decision making, etc., that are convenient for researchers but do not translate well to real-world applications. There is a need to study and document fairness problems and solutions in the wild.

In this paper, we report on a live deployment of a fairness-aware intervention on the website of the crowd-sourced microlending

platform Kiva.org. We discuss how the fairness concern was developed, defined, and measured; how the intervention was designed and deployed; and how the results were analyzed. The key finding from this study is that, rather than seeing a trade-off between the twin goals of recommendation accuracy and outcome fairness, the experiment found a degree of synergy: both fairness and our user engagement metrics increased. This provides evidence that, if properly integrated, we may be able to enhance important fairness measures within a system without a loss of overall accuracy or degradation of user experience.

Figure 1 shows the lender (user) interface on the Kiva.org website. Note that loans are displayed in multiple, themed rows or *slates* [11]. The first slate depicted is "Recommended" for one of this paper's authors; the second slate contains loans highlighted because of their impact; and the third slate shows loans that are close to completing their funding; other slates are available by scrolling down the page.

Each of these slates is generated according an independent logic and the interface as a whole forms a mixed hybrid recommender system [1]. Each slate is labeled with text that explains the logic behind it and encourages users to lend: for example, the description for the last slate says "Be the difference maker for these borrowers who only have a small amount remaining to be funded." A multi-armed bandit algorithm selects and orders the slates, making the interface itself is online and adaptive. Our intervention, as described below, is one that adds a fairness-specific slate to this system.

2 FAIRNESS-AWARE RECOMMENDATION

As with other forms of machine learning fairness, a fairness-aware recommender will focus on one of two fundamental concerns: individual or group fairness. Specifically, one can try to preserve the property that each individual is treated with equanimity on their relative merits or one can identify a protected group or groups and assert the goal of ensuring fair treatment for that group in the context of a larger population. However, these two views of fairness are often in direct competition with each other as noted in [8]. For example, giving preference to those with highest test scores may systematically disadvantage some group that, for irrelevant reasons, scores lower on that test. It is reasonable to expect that in any system, there may be a mix of multiple relevant fairness concerns that a system would like to honor, although most research in this area has concentrated on a single concern at a time [7].

In recommender systems, the typical two-view ontology of fairness is complicated by the multi-sided nature of many recommendation applications. The recommendations made by a music streaming site, for example, impact both the listeners (i.e., consumers of recommendation outputs) and also musical artists (i.e., providers of recommended material). Thus, we can speak of consumer-side and provider-side fairness concerns [2], and again, multiple concerns of both sides of the market may be active at any given time in a given application. In our application, we are primarily concerned with provider-side fairness: fairness towards borrowers.

Fairness-enhancing interventions can take a variety of forms. The survey by Ekstrand et al. [7], following the discussion in [14] and others, identifies three sites for intervention: pre-processing (e.g., re-sampling training data) that modifies system inputs, model enhancements (e.g., building fairness concerns into loss or optimization functions), or post-processing of recommendation outputs (e.g, re-ranking or inserting specific items into the output). Interestingly, the intervention described in this paper fits into none of these categories: it rather integrates fairness-oriented recommendations in manner of a "mixed" hybrid as defined in [1]: fairness-oriented results are presented in a separate slate added to the interface, and the recommendation outputs of other algorithms are left unchanged.

2.1 Fairness Metrics

Fairness is a complex and contested concept that has had numerous linguistic and technical definitions over a long history [10, 15]. As emphasized in [16], each context demands fairness considerations tailored to the task at hand. A variety of fairness constructs have been explored in relation to Kiva's lending context, including those noted in [18]. For the purposes of the work reported here, we use a notion of individual fairness relative to each loan, seeking to equalize the amount of time it takes for any particular loan to be funded: a goal of fairness with respect to funding time. However, to achieve this goal, we focus on a different key performance indicator (KPI), the "add to basket" (ATB) rate. This is the rate at which users who encounter a particular slate add an item from it to their online basket for potential funding. This is one step removed from the funding action. However, users' decisions to fund particular loans may be made for many reasons, and it can be difficult to disentangle these to isolate the impact of the recommender system. ATB rate, on the other hand, is directly associated with users' experience within the recommendation interface and so it is the metric that Kiva has used to understand the performance of its recommender system.

3 RESEARCH CONTEXT: KIVA MICROFUNDS

Kiva Microfunds uses crowdsourcing to provide access to capital for individuals and entrepreneurial groups, especially in the developing world, who are otherwise under-capitalized. Kiva partners with local organizations in countries across the globe and, as of 2022, has lent $1.74 billion to 4.3 million borrowers in 77 countries with the support of 2.1 million lenders[1]. Information about loan opportunities are listed on Kiva's site and promoted to its lenders (users), typically for thirty days or until the loan request is fully funded. Lenders can find loans to fund by browsing or searching the Kiva site. Kiva also promotes selected loan opportunities through a variety of means, including email advertisements and other promotions to users.

We focus here on the recommendation interface shown in Figure 1 where individual slates of loans associated with particular recommendation logics are shown to the user on Kiva's web site and through its mobile app. Each slate contains from 11 – 20 loans with 4 loans appearing in the initial presentation and additional entries available by paging right.

Some examples of slates include:

- Personalized: The "Recommended for ..." slate is constructed using a content-based recommender, which learns from the user's lending and activity history.

[1] https://www.kiva.org/about

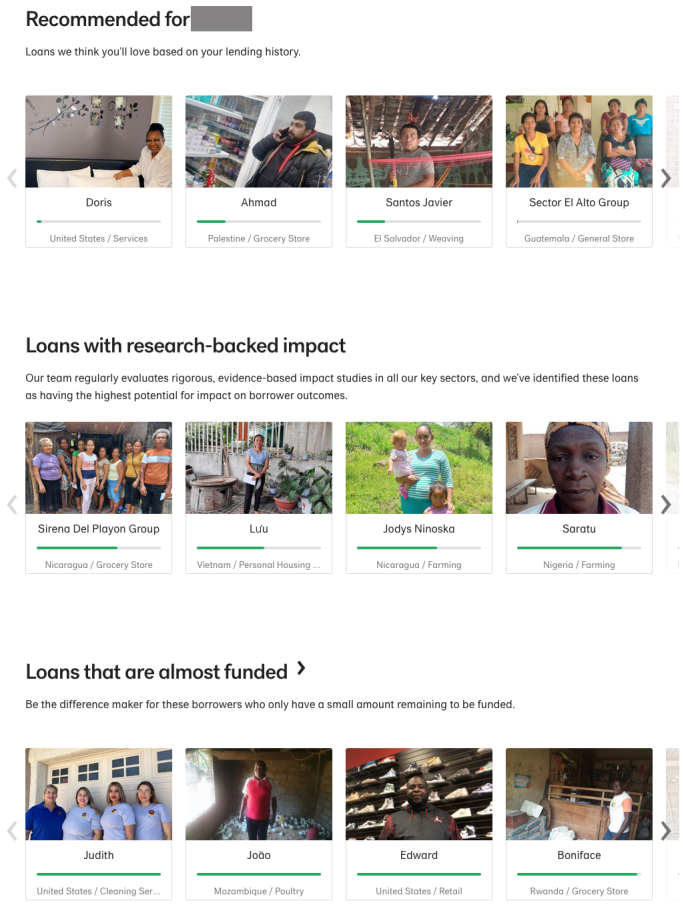


Figure 1: Example of a loan listing page on Kiva.org.

- Groups: Only contains loans that are from a collective or group of borrowers. These loans are often larger in amount, but can have significant economic benefits.
- Evidence-based Impact: Contains loans that are considered to have a high likelihood of positive economic impact, based on current research in international development.
- Countries you've lent to: Loans from countries prominent in the user's lending history.
- Loans almost funded: Loans that are close to their funding targets. Lenders are often motivated by the desire to put a borrower "over the top".

Except for the Personalized slate, the loans on each slate are ordered in terms of popularity, capturing trending activity on the site.

As noted above, the collection of slates presented to any given user is managed by an adaptive recommendation mechanism. Slate rankings are produced using Beta-Bernoulli bandits with Thomson sampling [4]. This technique has the benefit of adapting well to changes in rewards over time. Each of the approximately 20 possible slates is in effect a bandit itself, trying to optimize its reward independent of the other rows. The mechanism selects 10 slates from top to bottom, each selection being removed from the choices in the rows below. Based on user interactions with the presented slates (ATB events), the bandit updates its estimate of the current utility of each slate. Note that the bandit algorithm does not control the content of the slates, only what slates are presented and in what order. The Personalized slate consistently has the highest utility by this measure and usually appears in the first position.

Kiva's mission of improving global financial equity requires a focus on ensuring fairness in accessing capital and delivering loans. The recommendation system described here—directing users to certain loans on Kiva's platform—is vital to the success of its equity-promoting efforts; Kiva is thus a particularly compelling case study for the exploration of recommendation fairness. First, the fairness requirements at Kiva are driven by internal needs surrounding its philanthropic mission, rather than external demands, such as regulatory requirements that might be found in financial services or employment. A regulatory environment is more likely to provoke a defensive response related to fairness questions, which tends to hamper robust discussion of fairness properties in existing systems [6]. In addition, because of its philanthropic mission, it is reasonable to expect that Kiva's users will be receptive to fairness-oriented

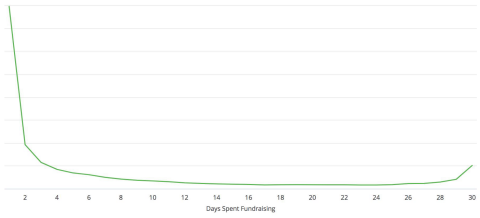


Figure 2: Distribution of funding time

interventions. Finally, as a hybrid organization, embodying the characteristics of a nonprofit along with the characteristics of a financial services institution, research embedded with Kiva is well-situated to have a broader and more generalizable impact across genres of institutions and sectors.

3.1 The "Time to Completion" Problem

Kiva's operational desire for fairness with respect to the rate of funding prompted an intervention in the area of recommending lending opportunities. The Kiva team observed that loans tend to receive more funding at the beginning and end of their lending period due to increased promotional focus on new loans and loans close to expiry. The beginning and ending nature of loan promotion created the situation where if a loan opportunity did not receive the requested amount quickly, it would remain incomplete until near the end of the lending window (if it received full funding at all).

Figure 2 clearly illustrates the beginning/ending phenomenon. It plots the distribution of funding times for loans in Kiva.org during a recent calendar year. A majority of loans are funded within the first four days of being listed. However, past that point, interest plateaus at a low level. Loans that are about to expire are attractive to many Kiva users and receive their own slate in the interface layout. Thus, the funding rate lifts slightly towards the end of the 30 day listing period.

Funding loans more quickly moves them out of the funding queue and enables Kiva's partner organizations to receive funds more quickly. This in turn increases the loan flow on the site and has the potential to greatly increase both the number and dollar quantity of loans being supported. Internal estimates suggested that increased promotion of low-probability loans could lead to an estimated additional 2000 loans being funded each year, supporting Kiva's goal of financial inclusion.

4 METHODOLOGY

In an effort to shorten time to completion—enabling more equitable access to capital for more loans—a Fairness slate labeled "Spotlighted by Kiva" was added to the lending interface of the Kiva site, highlighting borrowers whose loans were least likely to complete their funding on the current day. See Figure 3. Adding a separate slate was considered a relatively non-intrusive intervention for two reasons: first, this slate could be easily ignored by users who were not interested in what it is presenting in favor of other slates in the interface. Secondly, the contents of the other slates remained unchanged, so users familiar with the logic associated with these slates will to continue to get what they are expecting.

As noted above, the goal of the Fairness slate is to provide enhanced funding opportunities to loans in the plateau portion of their funding cycle. We operationalized this fairness concern as a requiring that we provide recommendation opportunities in inverse relation to the predicted funding of a loan at any given time. That is, we sought to promote the loans least likely to be fully funded on any given day. We can think of this problem as one of the predicting a loan's *failure to fund* (FTF) probability.

The amount of time that a loan has been on the site is strong predictor of its FTF likelihood. Therefore, to select the loans with the highest (worst) FTF score, we created a set of predictive models $\{P_1, \ldots, P_{30}\}$, corresponding to each day that a loan remains on the site for funding. P_{10}, for example, predicts whether a loan will fail to be fully funded on Day 10. Then for each loan, the system predicts whether the loan will fail to be funded today: P_1 is run on new loans, P_2 on loans that have been on the site one day, up to P_{30}. The predictors were built using an AutoML procedure [9] using 74 features covering loan metadata/information and measures of traffic, both raw (views) as well as normalized (views per session). In most cases, the AutoML mechanism chose gradient boosted decision trees as the most effective model.

Each day, every loan is scored using the P_i model appropriate to its time on the site, and the set of FTF scores is accumulated. Then a pool of 100 candidate loans is created using a random sample drawn using each loan's proportional FTF score (FTF / sum of FTF from all loans). The front-end then picks the highest-scoring 20 loans from this pool to display in the Fairness slate.

To test the new Fairness slate, we performed an A/B test over 20 days in 2021, creating a configuration in which the Control and Test conditions were identical in their interface presentation, except for the substitution of the Fairness slate for one of the other slates in the system. For the sessions chosen for testing, the layout bandit system was turned off so that the set of slates and their layout could be controlled. Instead, the overall ATB performance for each slate was used to generate a fixed ranking. To make our test a fairly strong one, we chose to insert the Fairness slate in the Test condition as the 2nd slate in the interface, after the Personalized slate. In this position, it replaced the Groups slate, which was the second strongest in terms of ATB. Our reasoning was that if the Fairness slate achieved ATB rates in this position comparable to the Groups slate, it would be working as well as anything short of the Personalized slate. It would demonstrate that our fairness-aware intervention was not detracting from the user experience or degrading the contribution of the recommender toward lending.

Prior study of order flow at Kiva demonstrated that ATB rates are highly differentiated between "logged-in" and "anonymous" users. Users who are logged in are much more likely to lend, and so their base ATB rates are much higher. Anonymous users may accumulate loans in their baskets and then log in or create an account to complete their transactions. However, many anonymous users are exploring the site and may not intend to loan at all, hence the smaller number of ATB events for these users. Grouping both sets of users together does not give a good picture of the impact on Anonymous users since their base ATB is so much smaller.

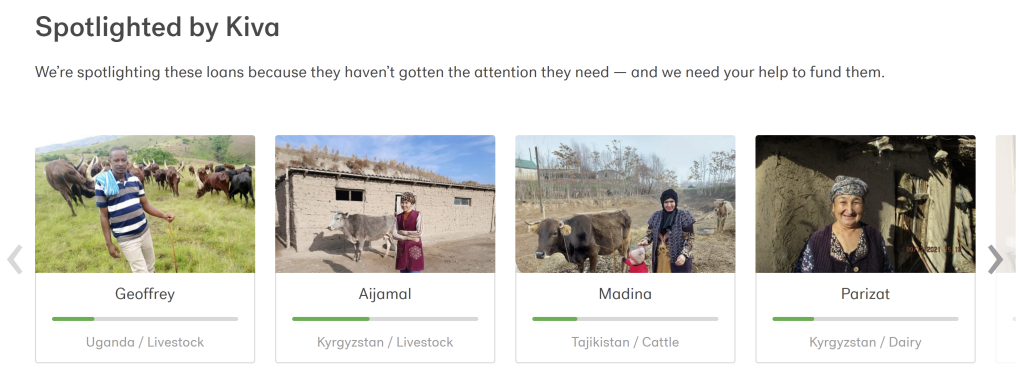


Figure 3: Example of a Spotlight row as shown on Kiva.org

We compared ATB rates between the Fairness slate in the Test group versus the Groups slate in Control group. We also examined ATB rates in the other 9 slates in the interface to see if the inclusion of the new slate impacted user activity elsewhere.

5 RESULTS

As shown in Table 1, for both the Logged In and Anonymous cases, the Spotlight slate attracted more ATB events, with a very large increase noted for Anonymous users. N is different in these cases; there were approximately 53% more sessions in the Anonymous set. All other slates showed no statistically-significant (at $p = 0.01$) differences in ATB events between Test and Control.

Table 1: Difference in mean "Add To Basket" (ATB) rate when the Fairness slate is displayed to a user instead of the Groups slate. Values statistically significant at $p = 0.01$. No other slate positions showed a significant difference between the conditions.

Condition	Difference from Control
Anonymous	126%
Logged-in	38%

Figure 4 shows the probability density on a normalized scale of the ATB frequency for Anonymous users (left) and Logged-in users (right). Figure 4a shows the Anonymous users where ATB events are considerably more rare in an absolute sense but the Fairness slate intervention had a larger overall impact. In Figure 4b we can see that Logged-in users are more likely to conduct transactions and so have a higher baseline ATB rate. Here the impact of the Fairness slate variant is smaller.

6 DISCUSSION

In this study, we see some of the complexities of implementing fair recommendation in practice, issues rarely addressed in the research literature. We note in particular the important constraint to be conservative in changes to an existing well-established interface and algorithm. Kiva opted not to modify one of its existing algorithms, for example, by adding constraints to enhance fairness; this might have the effect of discouraging current active users who have established expectations about how each slate is organized. Instead, we chose to add an additional fairness element to the already multi-slate interface.

The discourse around fairness in recommender systems often focuses on trade-offs between accurate recommendations and fair ones [17]. However, this research is largely based on experiments performed with historical data in which it is difficult to disentangle the effects of presentation bias from user preferences. Only a few published examples of fairness-aware interventions in live recommender systems platforms exist at this point; for example, [12, 13].

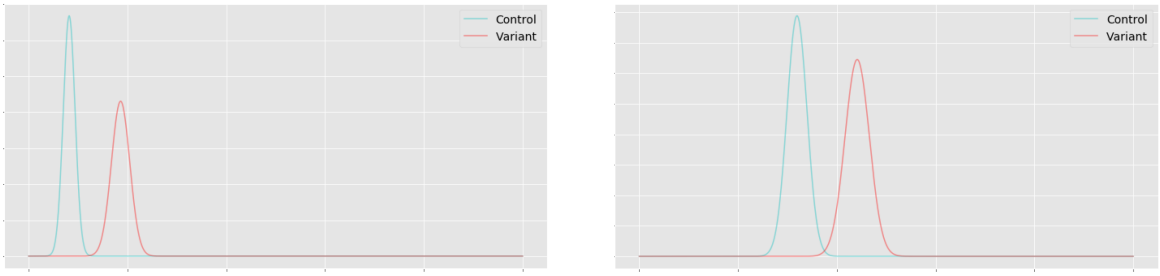
In contrast to the findings in [12], we did not find that user experience needs to be sacrificed to improve fairness, suggesting that the fairness / accuracy trade-off as currently conceived in the literature may be, at least in part, an artifact and not always present in real systems. In Kiva's context, both the business KPI (the "add to basket" events) and the fairness (the exposure of neglected loans) increased at the same time. The results of this A/B test were considered sufficiently positive that the Fairness slate was released as a permanent addition to the suite of options available to the layout bandit and this slate now regularly appears among the others shown in Kiva's web interface.

ACKNOWLEDGMENTS

Authors Burke, Voida, Kathait and Sonboli were supported by the National Science Foundation under grant awards IIS-1911025 and IIS-2107577. Nicholas Mattei was supported by NSF Grant IIS-2107505.

REFERENCES

[1] R. Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction* 12, 4 (2002), 331–370.
[2] Robin Burke. 2017. Multisided Fairness for Recommendation. *CoRR* abs/1707.00093 (2017). arXiv:1707.00093 http://arxiv.org/abs/1707.00093
[3] Robin Burke, Alexander Felfernig, and Mehmet H Göker. 2011. Recommender systems: An overview. *AI Magazine* 32, 3 (2011), 13–18.
[4] Olivier Chapelle and Lihong Li. 2011. An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (Eds.), Vol. 24. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2011/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf



(a) Anonymous Users **(b) Logged-in Users**

Figure 4: Normalized distribution of ATB values

[5] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

[6] Henriette Cramer, Jenn Wortman Vaughan, Ken Holstein, Hanna Wallach, Jean Garcia-Gathright, Hal Daumé III, Miroslav Dudík, and Sravana Reddy. 2019. Challenges of incorporating algorithmic fairness into industry practice. *FAT* Tutorial* (2019).

[7] Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz. 2021. Fairness and Discrimination in Information Access Systems. arXiv:2105.05779 [cs.IR]

[8] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (April 2021), 136–143. https://doi.org/10.1145/3433949

[9] Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems* 212 (2021), 10–66.

[10] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, 49–58.

[11] Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Tushar Chandra, and Craig Boutilier. 2019. SLATEQ: a tractable decomposition for reinforcement learning with recommendation sets. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI, 2592–2599.

[12] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the Conference on Information and Knowledge Management*. ACM, New York, 2243–2251.

[13] Rishabh Mehrotra, Niannan Xue, and Mounia Lalmas. 2020. Bandit based Optimization of Multiple Objectives on a Music Streaming Platform. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, New York, 3224–3233.

[14] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2020. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8 (Nov. 2020), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

[15] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This thing called fairness: disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36.

[16] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. ACM, New York, 59–68.

[17] Nasim Sonboli. 2022. *Controlling the Fairness / Accuracy Tradeoff in Recommender Systems.* Ph. D. Dissertation. University of Colorado, Boulder.

[18] Nasim Sonboli, Farzad Eskandanian, Robin Burke, Weiwen Liu, and Bamshad Mobasher. 2020. Opportunistic Multi-Aspect Fairness through Personalized Re-Ranking. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 239–247. https://doi.org/10.1145/3340631.3394846