Algorithmic Fairness, Institutional Logics, and Social Choice

Robin Burke[1] , **Amy Voida**[1] , **Nicholas Mattei**[2] and **Nasim Sonboli**[1]
[1]University of Colorado, Boulder
[2]Tulane University
{robin.burke, amy.voida}@colorado.edu, nsmattei@tulane.edu, nasim.sonboli@colorado.edu

Abstract

Fairness, in machine learning research, is often conceived as an exercise in constrained optimization, based on a predefined fairness metric. We argue that this abstract model of algorithmic fairness is a poor match for the real-world, in which applications are likely to be embedded within a larger context involving multiple classes of stakeholders as well as multiple social and technical systems. We may expect multiple, competing claims around fairness coming from various stakeholders, especially in applications oriented towards social good. We propose that computational social choice is a promising framework for the integration of multiple perspectives on system outcomes in fairness-aware systems and provide an example case of personalized recommendation for a non-profit.

1 Introduction

A substantial body of research on fairness in machine learning, especially in classification settings, has emerged in the past ten years, including formalizing various definitions of the concept of fairness [Chouldechova, 2017; Dwork *et al.*, 2012; Hardt *et al.*, 2016; Narayanan, 2018] and offering algorithmic techniques to mitigate unfairness under these definitions [Kamiran *et al.*, 2010; Pedreshi *et al.*, 2008; Zemel *et al.*, 2013; Zhang and Wu, 2017]. While many cases of problematic systems appear in popular literature, e.g., [O'Neil, 2016], only a small number of studies of deployed systems exist, e.g., [Chouldechova *et al.*, 2018; Mehrotra *et al.*, 2018; Beutel *et al.*, 2019b]. All too often, detailed studies of the impacts of these systems are hampered by their commercial nature, although there has been some recent sharing of this kind of experience [Holstein *et al.*, 2018; Cramer *et al.*, 2018].

One side effect of this lack of empirical grounding in real-world, deployed systems is that researchers tend to rely on highly simplified concepts of fairness in their metrics and algorithms. Generally, a single protected attribute and a single binary distinction define the problem. There is little recognition of the intersection of multiple fairness definitions and dimensions, although recent work has noted the benefits of combining multiple fairness definitions [Beutel *et al.*, 2019a]. Most existing research considers only a single protected class, and even in cases where multiple groups are considered, e.g., [Buolamwini and Gebru, 2018; Hebert-Johnson *et al.*, 2018; Kearns *et al.*, 2017; Zhu *et al.*, 2018], fairness is conceived using the same definition for all groups. Complex, multi-vocal notions of fairness arising from multiple stakeholders do not appear, in spite of decades of research in sociology and organizational studies demonstrating the complexity of such values in practice. Hence, we believe it is necessary to accommodate different definitions of fairness from different stakeholders, all of which must be integrated in a single framework. This nuanced understanding of the value of fairness is essential for capturing the richness of this social construct.

We are keenly interested in the application of algorithmic fairness in contexts where the requirements for fairness arise from an organization's mission, in contrast to a legalistic orientation, where the requirements for fairness take the form of legal/regulatory requirements imposed externally. A legalistic focus concentrates research effort on trying to ensure that a system will produce legally defensible results, based on particular standards. Although possibly influenced by regulatory regimes, mission-oriented fairness concerns are endogenous to an organization and thus the goal of fairness is held by internal stakeholders. We expect that such an orientation will be particularly important for non-profit organizations striving to achieve social good, but this is not the only application. Consider the work by Mehrotra *et al.* [2018] in which the music streaming service Spotify details its attempts to balance recommendation of highly-popular artists with the promotion of lesser-known names. There is no legal requirement for Spotify to promote less-popular artists, but to do so is beneficial to their platform ecosystem and their business model.

In accepting the complexity of fairness in mission-oriented contexts, we aim to widen the lens of algorithmic fairness towards the societal and organizational processes through which decisions are made and in which different groups contest for their own sense of what constitutes fair treatment. Social and political mechanisms decide what constitutes fairness, with results that differ widely by historical and social context. After all, student and senior citizen discounts constitute a form of ageist discrimination, but are widely considered acceptable in the United States. By extending our analysis back into this contested territory, we can render algorithmic fairness research more relevant to the contexts where it is most needed.

2 Institutional Logics and Values

Scholars in organizational studies widely concur that broader belief systems shape the way that stakeholders operate and make decisions within organizations, referred to as *institutional logics* [Thornton *et al.*, 2012]. Numerous studies have characterized the ways in which, in practice, *multiple* institutional logics are operating simultaneously within organizations—in competition and contestation with each other (e.g., [Besharov and Smith, 2014; Greenwood *et al.*, 2010; Pache and Santos, 2013; Reay and Hinings, 2009]). In the nonprofit context, for example, organizations are commonly required to negotiate the competing and often conflicting institutional logics that are part and parcel of being a mission-driven organization, e.g., prioritizing services to clients, and institutional logics derived from their often-public sector funders, such as fiscal efficiency and accountability (e.g., [Evers, 2005; Binder, 2007; Mullins, 2006]).

Connecting this body of research to the domain of computing, Voida *et al.* [2014] further found that technologies also embody institutional logics in the ways they instantiate particular values. Even when organizational stakeholders agree on the importance of a particular value, such as fairness, that value may not be operationalized in the deployed technology in a way that is harmonious with stakeholders' often-heterogeneous assumptions about and orientations towards how to put those values into practice. Such a mismatch can create significant challenges for organizations and the clients they serve. This research, then, suggests that fairness-aware recommendation systems will need to be able to harmonize across multiple, conflicting logics.

2.1 Example Application

Consider the loan recommendation situation of Kiva.org—a peer-to-peer micro-lending platform. The end users of Kiva are lenders who support entrepreneurs, typically from developing countries, by lending small amounts of money. The organization has the goal of providing equitable access to capital for all entrepreneurs who request loans, regardless of their geographic location, economic sector, gender, etc. This mission is instantiated through an online platform on which end-users search for and select entrepreneurs to fund. In this domain, a recommendation system that could promote entrepreneurs in certain underfunded areas, i.e., be more fair, would likely better serve the organization's mission of equity.

Yet, within Kiva, there are numerous classes of stakeholders, many working from different institutional logics and different—sometimes conflicting or competing—understandings of what equity means and/or how best to enable it in this complex, real world context. Fairness to an entrepreneur might mean that the quality of their business plan is privileged in the recommendation process. Fairness to the non-governmental organizations that help to serve as mentors and fiscal liaisons to the entrepreneurs might mean that each organizations' entrepreneurs are funded at equivalent rates. Fairness to Kiva's global mission might mean prioritizing funding entrepreneurs from systemically underfunded regions. And so on. In this context, then, one would need to conduct an analysis of the ways in which different stakeholders involved in the loan recommendation scenario understand fairness; and the development of any such system would need to consider how to prioritize and harmonize these multiple fairness concerns for the real-world system.

3 Social Choice and Fairness

As we have seen, fairness has a variety of definitions: reward, compensation, exogenous right, fitness [Moulin, 2004], and these different definitions may need to coexist in a given context. Our thumbnail sketch of fairness in Kiva.org above incorporates multiple such aspects. The borrower with a sound business plan deserves a *reward*; the borrower in an underdeveloped economic sector deserves extra promotion as *compensation* for lack of opportunity, etc. We should not assume that all fairness concerns are cut from the same cloth or can be measured in the same way.

Fairness has been a central concern of numerous technical and philosophical disciplines for centuries. Among these is the economic discipline of *social choice*, the study of how groups make decisions when each member is endowed with their own preferences [Arrow *et al.*, 2010]. To these considerations the field of computational social choice adds computational tools including algorithms, complexity, and big data [Brandt *et al.*, 2016] and incorporates research from multi-agent systems [Shoham and Leyton-Brown, 2008], looking at systems, the interactions between systems, and the behaviors of agents inhabiting those systems. These techniques have been successfully applied to a broad range of areas including markets on the internet [Moulin, 2018], routing data traffic [Kleinberg *et al.*, 1999], kidney exchange [Dickerson *et al.*, 2014], and employment screening [Schumann *et al.*, 2019].

We propose to re-conceptualize algorithmic fairness in mission-oriented contexts as an application of computational social choice and that doing so solves a number of key problems in algorithmic fairness. In a traditional social choice setting we have a finite set of agents $N = \{1, \ldots, n\}$ and a finite set of alternatives $A = \{1, \ldots, m\}$. Each agent $i \in N$ has a preference $\succsim_i$ over the alternatives. Typically these preference are expressed as a binary relation over the set A. Indeed, adopting a social choice perspective for both matching/allocation and voting have started to appear in the fairness literature including for finding fair matches in ride-sharing [Sühr *et al.*, 2019], finding fair group recommendations through viewing them as elections [Chakraborty *et al.*, 2019], and most closely to our setting, building a fair recommendation system through viewing it as matching market [Patro *et al.*, 2020]. However, none of this research considers multiple fairness concerns on the provider side as required in the Kiva case.

In our algorithmic fairness scenario as shown in Figure 1, the agents are not actors in a traditional sense (like individual voters in a democracy) but rather *fairness concerns* that emerge from institutional logics within an organization. The alternatives over which these fairness concerns contend will differ depending on the organization, what resources it has, and how these resources may be allocated. The only capability agents need is the ability to compute their preferences over decision outcomes in the classic social choice sense. Combining those preferences is the task of a social choice function in

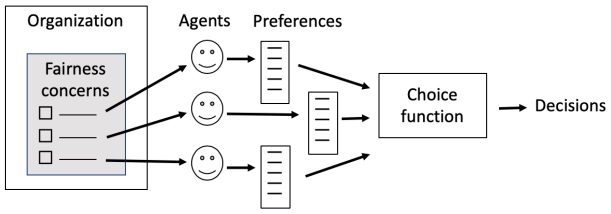

Figure 1: Social choice fairness framework

the figure that produces a final result.

For Kiva, the fairness concerns may center around different aspects of the loans that may be recommended: the associated entrepreneurs, the sectors of the economy in which they work, and other considerations. The point of decision is the moment at which recommendations must be delivered to an end-user. Within this decision, the system is ranking different loans to present and it is therefore natural to think of loans being ordered with respect to different preferences, including the end-user's own personalized interests.

However, the formulation we propose is not limited to personalized recommendation and to microlending. Consider a classic fairness-aware machine learning task, the college admissions scenario studied in [Friedler *et al.*, 2016][1]. The concerns for different aspects of diversity or fairness may be considered as having preferences over which set of students should receive an "admit" label in the classification operation. A fair admissions process is one in which the interplay between these different concerns resolves to a final outcome representing, as best as possible, the collective preference so expressed. Here we can leverage research from the preference handling and computational social choice communities on preference formalisms and compact representations to address issues in representing these complex, combinatorial concerns [Mattei and Walsh, 2017].

Conceptualizing multi-aspect fairness as social choice does not mean that algorithmic fairness will suddenly be solved. Many social choice problems are known to be NP-hard or to have no solutions when even reasonable constraints are imposed. For example, using social choice mechanisms like multi-winner voting to find proportional rankings is computationally hard [Skowron *et al.*, 2017]. Researchers who adopt this perspective will still have plenty of problems to keep them occupied. However, what this move does offer is a more natural way to derive and express fairness concerns, an expansive formalism for concerns defined in different ways, and a better way to explain the operation of such fairness-aware systems using, e.g., axiomatic analysis of the choice rules.

4 Non-deterministic Ranking

As an example of the kind of solution that a social choice perspective on algorithmic fairness makes possible, we outline here an approach to implementing fairness in personalized recommender systems. Recognizing the tension between group and individual fairness [Friedler *et al.*, 2016], we will be assuming a setting in which group fairness / non-discrimination is sought. We assume that the fairness concerns consist of groups over which recommendation results should be made fair, according to some fairness metric, and under some metric of recommendation outcomes. We also assume that the system is delivering a large number of recommendation outcomes over time. Thus, our objective is not that each individual list meets some fairness target but rather than our system can be fair in expectation over some time period.

We can think of this as a repeated choice environment, which lends itself to the use of probabilistic social choice methods, which have the advantage of greater tractability [Brandl *et al.*, 2016]. Because all of our fairness concerns are derived through a deliberative process within one organization and should, in principle, be aspects of the organizational mission, strategic aspects of preference disclosure are less significant than in other social choice contexts [Conitzer *et al.*, 2007]. There is a vast literature on repeated social choice and many algorithmic options, depending on specific problem characteristics. As an example, consider the simple non-deterministic mechanism in the *random serial dictator* model: the algorithm chooses randomly and with equal probability among the agents (in our case each fairness concern) and the chosen agent gets to impose their preferred ranking on the outcome [Brandt *et al.*, 2016].

5 Conclusion

We have argued for an approach to algorithmic fairness that does not take fairness as an externality but rather assumes that fairness concerns arise from business models, organizational missions, and stakeholder diversity. It is natural, therefore, to expect multiple institutional logics to be operative and multiple fairness concerns to arise. Thus we need a flexible and general characterization of fairness objectives and a way to allow multiple such objectives to interact in deriving outcomes.

The field of computational social choice provides an avenue for re-conceptualizing algorithmic fairness in a way that foregrounds the multiple and contested definitions likely to arise in practice. Social choice and welfare economics more generally have a long history of grappling with and reasoning about problems of fairness, and with balancing the concerns and preferences of multiple groups.

We have shown that a natural formulation of algorithmic fairness is to represent fairness concerns as actors with preferences over system outcomes. As a benefit, some of the thorny issues of social choice are ameliorated in this setting since the number of such concerns will be tractably small and, as noted above, the incentives for strategic "gaming" of the algorithm minimal. We anticipate that this formulation of algorithmic fairness will offer rich opportunities for research and system development in both recommender systems and social choice.

Acknowledgments

Authors Burke and Sonboli were supported in part by the National Science Foundation under Grant No. 1911025.

[1]The Borda voting rule in social choice, i.e., the derivation and combination of scores, is precisely the methodology most commonly used in college admissions to rank students based on different aspects of their backgrounds [Kretchmar, 2006].

References

[Arrow *et al.*, 2010] Kenneth J Arrow, Amartya Sen, and Kotaro Suzumura. *Handbook of social choice and welfare*, volume 2. Elsevier, 2010.

[Besharov and Smith, 2014] Marya L Besharov and Wendy K Smith. Multiple institutional logics in organizations: Explaining their varied nature and implications. *Academy of management review*, 39(3):364–381, 2014.

[Beutel *et al.*, 2019a] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2212–2220, 2019.

[Beutel *et al.*, 2019b] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 453–459, 2019.

[Binder, 2007] Amy Binder. For love and money: Organizations' creative responses to multiple environmental logics. *Theory and society*, 36(6):547–571, 2007.

[Brandl *et al.*, 2016] Florian Brandl, Felix Brandt, and Hans Georg Seedig. Consistent probabilistic social choice. *Econometrica*, 84(5):1839–1880, 2016.

[Brandt *et al.*, 2016] F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, editors. *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[Chakraborty *et al.*, 2019] Abhijnan Chakraborty, Gourab K Patro, Niloy Ganguly, Krishna P Gummadi, and Patrick Loiseau. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 129–138, 2019.

[Chouldechova *et al.*, 2018] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

[Chouldechova, 2017] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[Conitzer *et al.*, 2007] Vincent Conitzer, Tuomas Sandholm, and Jérôme Lang. When are elections with few candidates hard to manipulate? *Journal of the ACM (JACM)*, 54(3), 2007.

[Cramer *et al.*, 2018] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. Assessing and addressing algorithmic bias in practice. *interactions*, 25(6):58–63, 2018.

[Dickerson *et al.*, 2014] John P. Dickerson, Ariel D. Procaccia, and Tuomas Sandholm. Price of fairness in kidney exchange. In *International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1013–1020. IFAAMAS/ACM, 2014.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[Evers, 2005] Adalbert Evers. Mixed welfare systems and hybrid organizations: Changes in the governance and provision of social services. *Intl Journal of Public Administration*, 28(9-10):737–748, 2005.

[Friedler *et al.*, 2016] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[Greenwood *et al.*, 2010] Royston Greenwood, Amalia Magán Díaz, Stan Xiao Li, and José Céspedes Lorente. The multiplicity of institutional logics and the heterogeneity of organizational responses. *Organization science*, 21(2):521–539, 2010.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[Hebert-Johnson *et al.*, 2018] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948, 2018.

[Holstein *et al.*, 2018] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, and Hanna M. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? *CoRR*, abs/1812.05239, 2018.

[Kamiran *et al.*, 2010] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 869–874, University of Technology Sydney, Australia, 2010. IEEE.

[Kearns *et al.*, 2017] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

[Kleinberg *et al.*, 1999] Jon Kleinberg, Yuval Rabani, and Éva Tardos. Fairness in routing and load balancing. In *40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 568–578. IEEE, 1999.

[Kretchmar, 2006] Jennifer Kretchmar. Assessing the reliability of ratings used in undergraduate admission decisions. *Journal of College Admission*, 192:10–15, 2006.

[Mattei and Walsh, 2017] Nicholas Mattei and Toby Walsh. A PREFLIB.ORG Retrospective: Lessons Learned and New Directions. In U. Endriss, editor, *Trends in Computational Social Choice*, chapter 15, pages 289–309. AI Access Foundation, 2017.

[Mehrotra *et al.*, 2018] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*, pages 2243–2251, 2018.

[Moulin, 2004] Hervé Moulin. *Fair division and collective welfare*. MIT press, 2004.

[Moulin, 2018] Hervé Moulin. Fair division in the age of internet. *Annual Review of Economics*, 2018.

[Mullins, 2006] David Mullins. Competing institutional logics? local accountability and scale and efficiency in an expanding nonprofit housing sector. *Public Policy and Administration*, 21(3):6–24, 2006.

[Narayanan, 2018] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, 2018.

[O'Neil, 2016] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[Pache and Santos, 2013] Anne-Claire Pache and Filipe Santos. Embedded in hybrid contexts: How individuals in organizations respond to competing institutional logics. In *Institutional logics in action, part B*, pages 3–35. Emerald Group Publishing Limited, 2013.

[Patro *et al.*, 2020] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of The Web Conference 2020*, pages 1194–1204, 2020.

[Pedreshi *et al.*, 2008] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, New York, NY, USA, 2008. ACM.

[Reay and Hinings, 2009] Trish Reay and C Robert Hinings. Managing the rivalry of competing institutional logics. *Organization studies*, 30(6):629–652, 2009.

[Schumann *et al.*, 2019] Candice Schumann, Zhi Lang, Nicholas Mattei, and John P. Dickerson. Group fairness in bandit arm selection. *CoRR*, abs/1912.03802, 2019.

[Shoham and Leyton-Brown, 2008] Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations*. Cambridge University Press, 2008.

[Skowron *et al.*, 2017] Piotr Skowron, Martin Lackner, Markus Brill, Dominik Peters, and Edith Elkind. Proportional rankings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 409–415, 2017.

[Sühr *et al.*, 2019] Tom Sühr, Asia J Biega, Meike Zehlike, Krishna P Gummadi, and Abhijnan Chakraborty. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3082–3092, 2019.

[Thornton *et al.*, 2012] Patricia H. Thornton, William Ocasio, and Michael Lounsbury. *The institutional logics perspective: A new approach to culture, structure, and process*. Oxford University Press on Demand, 2012.

[Voida *et al.*, 2014] Amy Voida, Lynn Dombrowski, Gillian R. Hayes, and Melissa Mazmanian. Shared Values/Conflicting Logics: Working Around e-Government Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3583–3592, New York, NY, USA, 2014. ACM.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

[Zhang and Wu, 2017] Lu Zhang and Xintao Wu. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics*, 4(1):1–16, 2017.

[Zhu *et al.*, 2018] Ziwei Zhu, Xia Hu, and James Caverlee. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1153–1162, 2018.