We Need Fairness and Explainability in Algorithmic Hiring

Blue Sky Ideas Track

Candice Schumann
University of Maryland
College Park, Maryland, USA
schumann@cs.umd.edu

Jeffrey S. Foster
Tufts University
Medford, Massachusetts, USA
jfoster@cs.tufts.edu

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

John P. Dickerson
University of Maryland
College Park, Maryland, USA
john@cs.umd.edu

ABSTRACT

Algorithms and machine learning models, including the decisions made by these models, are becoming ubiquitous in our daily life, including hiring. We make no value judgment regarding this development; rather, we simply acknowledge that it is quickly becoming reality that automation plays a role in hiring. Increasingly, these technologies are used in all of the small decisions that make up the modern hiring pipeline: from which resumes get selected for a first screen to who gets an on site interview. Thus, these algorithms and models may potentially amplify bias and (un)fairness issues for many historically marginalized groups. While there is a rapidly expanding literature on algorithmic decision making and fairness, there has been limited work on fairness specifically for online, multi-stakeholder decision making processes such as those found in hiring. We outline broad challenges including formulating definitions for fair treatment and fair outcomes in hiring, and incorporating these definitions into the algorithms and processes that constitute the modern hiring pipeline. We see the AAMAS community as uniquely positioned to address these challenges.

KEYWORDS

Blue Sky; Algorithmic Fairness; Bandits

ACM Reference Format:
Candice Schumann, Jeffrey S. Foster, Nicholas Mattei, and John P. Dickerson. 2020. We Need Fairness and Explainability in Algorithmic Hiring. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 5 pages.

1 INTRODUCTION

"Hiring is rarely a single decision point, but rather a cumulative series of small decisions." So begins a recent report on *automated hiring processes* released by the non-profit group UpTurn [14], before recommending that digital sourcing firms begin explicitly addressing concerns of fairness and bias at every step of the hiring process. Indeed, at various decision points in the hiring process, algorithms already determine who sees which job advertisements; estimate the expected performance of an applicant; select which applicants to screen more heavily and with whom to match them; and forecast salary and other benefits necessary to ensure a successful offer. Thus, issues of bias or fairness at one stage of this procedure may lead to unexpected or amplified issues at a later stage of the process.

In addition to the difficulty of these decisions on their own, there are a number of regulatory and legal requirements that must be met at each stage of the hiring process. As a recent Facebook settlement showed, the tools, platforms, and techniques developed to streamline hiring can be subtly—or blatantly—illegal [24]. These requirements are complicated by the presence of multiple stakeholders: governmental regulators, hiring managers, employees, line managers, and myriad others involved in hiring and employment.

While one can argue that we may not need algorithmic hiring, the fact is that platforms and websites such as LinkedIn, ZipRecruiter, and Indeed are making these tools available to businesses of any size, and that large businesses are experimenting or have experimented with automated hiring techniques [25]. Thus, algorithmic processes are being deployed in the real-world, and it is incumbent on computer science researchers to ensure that the algorithms we create are aware of both fairness and legal compliance for these processes. There is already ample evidence from the areas of lending and pre-trial detention (bail) and policing that the algorithms that are deployed can have significant, and sometimes harmful, impacts on individuals lives [21]. There is a need for novel techniques from data science, artificial intelligence, and machine learning to ensure our algorithms act within the constraints set forth by business process, laws, social norms, and ethical guidelines [41].

One shortcoming of current research into algorithmic fairness is its focus on a *single* decision point [21]. As depicted in Figure 1, modern hiring is rarely a single step process [14]. It is the culmination of a series of steps, much like pre-trial detention and other decisions of consequence, and we currently lack the algorithmic tools and techniques to adequately address this challenge. Techniques developed to address these challenges can also be applied to many settings where we have a "prioritization funnel" setting, such as customer acquisition or government sourcing.

We argue for concentrated research around the thesis that:

> *Data-driven approaches to measuring and promoting fairness and explainability to each of the concerned stakeholders at a single stage of the hiring process can*

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

be extended—in a principled way—to the full, multi-stage hiring process.

It is important to note that the application of research in this area will not just be in the hiring scenario. The techniques developed here, along with a number of results in peer evaluation [5] and other areas of social choice, including matching [15], will enable the creation of algorithmic tools that are both fair and efficient. These tools can and should be deployed in any situation where we are attempting to select a set of candidates (or items, or interventions) from a large pool or allocate other scarce resources, subject to various constraints over the selection and reviewing process [41]. These technologies could be applied to internal product ideation and review [53], academic proposal reviewing [27], advertisement/campaign selection [34], or indeed any setting where we need to collect recommendations over a large set from experts.

We detail the limits of current research into fairness and its shortcomings with respect to the challenge of algorithmic hiring. We detail both past and current work at AAMAS that demonstrates the communities potential impact in the area. Finally, we close with additional ideas we see as research directions for the community.

2 FAIRNESS IN ONLINE, MULTI-STAGE DECISION-MAKING ALGORITHMS

Within computer science, economics, and operations research circles many of the problems that are encountered in hiring are typically modeled in the *multi-armed bandit (MAB)* setting [48]. Indeed, bandit-based algorithms have received significant attention in the literature for their use in content recommendation [33], advertising, and hiring [13, 43]. Additionally, bandit algorithm development, and reinforcement learning in general, is a core topic at AAMAS with multiple sessions every year.

In the basic MAB setting, there are *n arms*, each associated with a fixed but unknown reward probability distribution [4, 32]. At each time step $t \in [T]$, an agent pulls an arm and receives a reward that is independent of any previous action and follows the selected arm's probability distribution independent of the previous actions. The goal of the agent is to maximize the collected reward over all timesteps. A generalization of MAB is the contextual multi-armed bandit (CMAB) problem where the agent observes a d-dimensional *context*, to use along with the observed distribution of rewards of the arms played, in order to choose a new arm. In the CMAB problem, the agent learns the relationship between contexts and rewards and select the best arm [3].

Examples of practical applications of MAB algorithms include algorithms for selecting what advertisements to display to users on a webpage [35], systems for dynamic pricing [36], and content recommendation services [33]. Indeed, such ML-based decision-making systems continue to expand in scope, making ever more important decisions in our lives such as setting bail [21], making hiring decisions [13, 43], and policing [42]. Thus, the study of the properties of these algorithms is of paramount importance [19].

Yet, the use of MAB-based systems often results in behavior that is societally repugnant. Sweeney [49] noted that queries for public records on Google resulted in different associated contextual advertisements based on whether the query target had a traditionally African American or Caucasian name; in the former case, advertisements were more likely to contain text relating to criminal incidents. Similar instances continue to be observed, both in the bandit setting and in the more general machine learning world [38]. In lockstep, the academic community has begun developing approaches to tackling issues of (un)fairness in a variety of settings.

Recently, a Computing Community Consortium whitepaper on fairness research specifically identified that most studies of fairness are focused on classification problems [19]. These works define a statistical notion of fairness, typically a Rawlsian notion of equal treatment of equals [40], and seek to constrain algorithms to abide by these constraints. Two fundamental issues identified by Chouldechova and Roth [19] that we believe are unaddressed by the current literature, and that we call on the AAMAS community to address, are extensions to notions of *group fairness* and looking at fairness in *online, dynamic systems*, e.g., the contextual bandit setting. We envision the research community addressing these gaps by formalizing and providing algorithms for definitions of fairness and bias.

The recent restructuring of the AAMAS into themes makes the areas that can contribute to these topics of fairness to groups of agents and online problems all the more clear. We see the following areas specifically as both sources of ideas and nexuses for collaboration around fairness in sequential decision making.

Area 7 – Markets / Game Theory. Mechanism and market designers are both interested in fairness towards the agents that participate. We see the game theory community within AAMAS as being particularly helpful when it comes to analyzing the incentives at play among classes of stakeholders in the hiring process, e.g., competing firms, or a single firm and a single candidate, or hiring managers within a firm.

Area 6 – Learning and Adaptation. Multi-armed bandits and other forms of reinforcement and online learning have been core to AAMAS since its inception [48]. Indeed, there have been numerous MAB papers at AAMAS recently that also deal with humans/crowdsourcing [39], fairness and diversity [43], and/or incorporating biased human feedback [50], to name just a few. Hence, we feel that the AAMAS community is able to help with this core topic.

Area 1 – Coordination, Organizations, and Norms. Many of the algorithmic hiring systems are both learning *from* and interacting *with* multiple stakeholders including hiring managers, line managers, and employees, in real time. The systems are making decisions in environments with multiple competing interests. Much like Area 7, researchers in these areas will be key in advancing this overall agenda. Furthermore, we believe research into multi-stage fairness could more closely tie together Areas 1, 6, and 7.

3 FAIRNESS IN THE HIRING PROCESS

The pipeline of a typical algorithmic hiring process is depicted in Figure 1. In this process, a set of applications is screened by either humans, algorithms, or a combination of both. After this initial screening and selection, applications are scored/ranked and many are discarded. After this an iterative process of allocating resources, e.g., requests for additional documentation; online or in-person interviews; and group discussion are committed to refine the initial

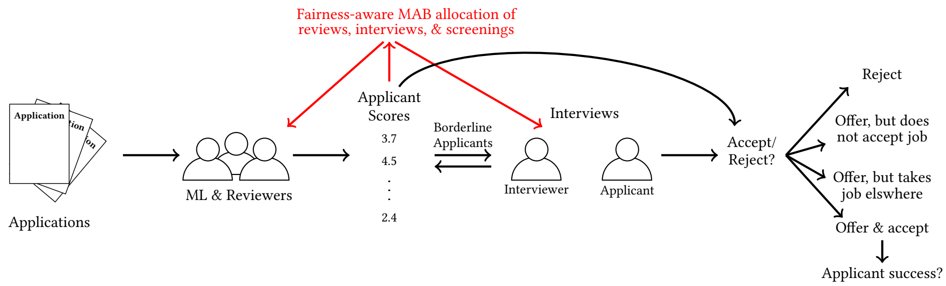

Figure 1: A sample current tiered hiring process (in black) and interventions proposed by this blue sky submission (in red).

ranking. After this, offers and/or rejections are sent to one or more candidates from the pool and the candidate provides a response.

We are proposing a focused research plan into a data-driven decision support process that draws inferences in part based on observed and estimated features of humans—and such tools are increasingly known to result in unexpected or adverse impact on dimensions such as fairness and bias [38]. We acknowledge that both our and others' initial work in this space, as well as our proposed extension to the more realistic multi-stage selection setting, may exacerbate issues of *fairness*. Thus, we also propose to incorporate recent definitions of fairness from the machine learning community into our tiered model. Such definitions do not fully capture the needs or wants of practitioners [28]; yet, we believe developing systems that are amenable to general definitions of fairness will be useful, because those definitions are evolving, and will continue to evolve, over time. In our exploratory work, we adopt a subset of the standard notions of fairness, and we perform analysis on real admissions data [43, 44]; still, much work remains to align systems to be fielded with the aggregate preferences of stakeholders.

It is important to ensure that the entire pipeline is capable of recognizing *fair treatment* and/or *fair outcome* (and possibly others) in the multi-armed bandit setting [30]. When modeling the hiring process as a MAB problem we have a set of arms $a \in A$, such that each applicant is an arm a, and where A is partitioned into L groups $A = P_1 \cup P_2 \cup \cdots \cup P_L$, but now corresponding to specific sensitive attribute groups. These attributes could represent self-reported gender, race, and country of origin. We have begun work in this direction, with preliminary results appearing at a NeurIPS-19 workshop [45]. We re-emphasize that, throughout, our models will be built to accept a host of fairness and parity measures; still, it is important to provide concrete plans for specific definitions of each.

We note that notions of "fairness," "bias," and "explainability" are (i) definable in many ways [21] and (ii) necessarily different based on application areas, societal norms, and policy-maker preferences. However, in hiring, credit, and housing there are a number of federally protected features that one must not use in the decision making process and also must not use for explanation. Simply removing these features from consideration by our algorithms is not enough, and we must actively ensure the fairness criteria is enforced across these features [17]. Thus, we endeavor to remain somewhat definition-agnostic in our modeling work, and then explicitly instantiate a definition when needed (e.g., we plan to use the well-known *equality of opportunity* [26] definition of fairness in our earliest experiments). However, our proposed approaches should generalize to a whole host of fairness or parity measures, so long as the measure of bias/fairness can be written as a linear constraint on conditional moments of predicted distributions over predictions, ground truth, and protected attributes [2].

A closely related area to our work is the research into fairness in rankings [46], multi-stakeholder recommender systems [1], and item allocation [9, 10]. When algorithms return rankings for an individual to select from one must pay attention to the ordering and the positioning of various groups [46]. One can see this as an application of the group fairness concept to the slates that are chosen for display. A particular aspect of recommendation systems that one needs to keep in mind is that often there are different stakeholders: the person receiving the recommendation, the company giving the recommendation, and the businesses that are the subjects of recommendation [1]. Finally, when goods are allocated, such as housing or subsidies one may need to observe both individual and group fairness [10]. Indeed, group fairness is specifically important in, e.g., Singapore, which has specifically enforced notions of group fairness when allocating public housing [9].

4 A FIRST STEP: AN INITIAL FRAMEWORK TO MODEL "FAIR" TIERED HIRING

We propose a multidimensional approach to tackling issues in the efficient and fair *gathering* and *aggregation* of information by hiring managers, which jointly compose part of a decision support system for potential job offer *decisions*. We propose to use the concept of *structured interviews* [16, 51], used widely in industry as well as in some academic programs (e.g., Fisk-Vanderbilt [47]); then, drawing on prior work [44], we will cast *tiered* hiring as a combinatorial pure exploration (CPE) problem in the stochastic multi-armed bandit setting [18]. The goal is to select a cohort of applicants after narrowing the pool after successive stages or tiers. Each tier or interviewing stage has an associated strength of arm pull, similar to (indeed, generalizing the concept of) the weak and strong arm pulls introduced in prior work [43]. The strength determines the confidence of the signal generated by the reviewer/interview as well as the cost of performing an arm pull; allowing us to generalize to more than two types of arm pulls, increasing usefulness.

Specifically extending this idea to a **multi-tiered setting**, e.g., in the graduate admissions case, the process could include an initial screening for minimum qualifications such as GPA, then an application review, followed by a Skype interview, and finally an in-person interview. Figure 1 gives an example tiered hiring process, and shows (in red text) where our proposed interventions fit into the present hiring system. In the tiered setting each stage creates a short list and the applicant pool is narrowed. In other words, during each stage K_i arms (i.e., applicants) move on to the next stage (i.e., we remove $K_{i-1} - K_i$ arms), where $n = K_0 > K_1 > \cdots > K_{m-1} > K_m = K$). Therefore, each stage i could be considered a cohort selection problem where K_i applicants need to be selected in order to maximize some objective function. Others have proven results under a linear objective, as is standard in the Top-K MAB literature [44]; general results would be useful.

The presently-developed methods allow for the promotion of diversity in the final cohort of applicants (e.g., graduate students). Dovetailing with this, the **fairness of the review process** is also important. In the MAB setting, we propose that the AAMAS community build on work in incorporating constraints into the MAB framework [7] and extend this work with methods from the fairness in machine learning literature [8, 20] such as those developed within the silos of fairness of treatment and fairness of outcome. Of particular value would be merging these criteria into the single-level and multi-tiered settings, exploring theoretical metrics such as the impact on overall economic efficiency due to the use of a "fair" objective, and experimental validation on sensitive attributes such as self-reported gender, race, and country of origin that are available in our real data sets.

5 BLUE SKY RESEARCH CHALLENGES

As noted earlier, we are *not* making a value judgment regarding the use of automated systems in hiring; rather, we note that this is, increasingly, reality. We are also *not* making value judgments regarding particular definitions of fairness and/or bias in machine learning. Our goal here is to develop *general* and *principled* systems for tiered hiring that can incorporate *many* definitions of fairness.

We are working on extending our current research to incorporate different notions of fairness that could be deployed on a number of already-fielded MAB-based systems [45]. We plan to extend these definitions to a tiered model [44] and investigate theoretically the "price of fairness" [11] in these systems. This initial work may close the gap on a single point (the hiring), but there is still much work to be done. Some of our initial research has addressed questions of transparency, constraints, and fairness when working with multi-armed bandit algorithms [6, 7, 30, 43, 45]. Yet, these are small steps taken toward a larger research goal. We see the following issues as still omnipresent concerns, ripe for work by researchers from the AAMAS community.

(1) How should we allocate effort—e.g., budget, interview slots—along the hiring pipeline? While we have begun to address this gap there are still challenges that remain. Included in these challenges is maintaining notions of diversity at every stage of the pipeline, and not just at discrete points.
(2) How can we explain the decisions made by the complete algorithmic process in a transparent and compliant way? With (inter-)national regulation like the newly-established GDPR [52] and the right to object and right to rectification, we need to build pipelines for decisions that are not only fair but capable of being audited.
(3) How can we incorporate fairness into other automated screening tools that we are beginning to see? For instance, chatbots are starting to be used to gather pre-interview data with clients and the need to address concerns around usability and access are almost completely untouched.
(4) How do we choose the features to select when building models for hiring? Which features are predictive, which are not, and which are protected? While the UpTurn study [13] states that employers should disclose all relevant features, the selection of these features is a ethically-laden decision. While there has been recent work in this area [37] further exploration is necessary.
(5) There has been extensive recent work in budget-limited and other constrained bandit models including limiting rounds [54], policy thresholds [55], and unknown, budget constrained cost distributions [23]. Exploring models with resource and budget constraints necessary for the hiring process is an important direction.
(6) So far, we have assumed individuals have fixed group membership and that these group memberships do not overlap. Generalizing fairness definitions to work for intersectional fairness and settings where memberships in protected groups may change at every timestep t would fit more real world applications. One step forward might leverage results from work on bandits with non-stationary rewards [12]. Additionally, other group fairness definitions such as Equalized Opportunity should be converted to the MAB setting [26].
(7) Algorithmic transparency to the end user is important, as discussed, but equally important is maintaining human involvement in the training, validation, and deployment process. We conjecture (and sincerely hope!) that no hiring process will become entirely automated—so we must ensure that the algorithms and systems we build are capable of working with, potentially biased, human input at every stage.
(8) In our previous work [43, 44] we explored an objective that balances both individual utility and the diversity of the set of arms returned. Research has shown that a more diverse workforce produces better products and increases productivity [22, 29]. Thus, such an objective is of interest to our application of hiring workers. Note that diversity, while related, is distinct from fairness. Trying to balance both diversity and fairness should be looked at more deeply since both diversity and fairness are important in the hiring process.
(9) We need a new definition of *fair outcomes for the MAB setting*. Typically, equality of opportunity fairness is used in classification tasks. We can formulate a strict definition of equal opportunity for bandits, but a hard constraint may be too strict a definition, or may not align with the expressed preferences of stakeholders. Instead, it may be necessary to define notions of fairness that straddle the line between individual and (sub-)community [31]. And, indeed, it may be necessary to balance notions of fairness and economic efficiency across both sides of the market, so as to promote truthful participation of both firms and workers in this ubiquitous and increasingly automated process.

Acknowledgments. Schumann and Dickerson were supported by NSF IIS RI CAREER #1846237 and a Google Faculty Research Award.

REFERENCES

[1] Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder Recommendation and its Connection to Multi-sided Fairness. *arXiv preprint arXiv:1907.13158* (2019).

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning (ICML)*. 60–69.

[3] Shipra Agrawal and Navin Goyal. 2013. Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International Conference on Machine Learning (ICML)*. 127–135.

[4] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47, 2-3 (2002), 235–256.

[5] H. Aziz, O. Lev, N. Mattei, J. S. Rosenschein, and T. Walsh. 2018. Strategyproof Peer Selection using Randomization, Partitioning, and Apportionment. *Artificial Intelligence Journal* (2018).

[6] A. Balakrishnan, D. Bouneffouf, N. Mattei, and F. Rossi. 2018. Using Contextual Bandits with Behavioral Constraints for Constrained Online Movie Recommendation. In *Proc. of the 27th Intl. Joint Conference on AI (IJCAI)*.

[7] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. 2019. Incorporating behavioral constraints in online AI systems. In *AAAI Conference on Artificial Intelligence*, Vol. 33. 3–11.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. Fairness and Machine Learning. fairmlbook. org, 2018. *URL: http://www. fairmlbook. org* (2018).

[9] Nawal Benabbou, Mithun Chakraborty, Edith Elkind, and Yair Zick. 2019. Fairness Towards Groups of Agents in the Allocation of Indivisible Items. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 95–101.

[10] Nawal Benabbou, Mithun Chakraborty, Xuan-Vinh Ho, Jakub Sliwinski, and Yair Zick. 2018. Diversity constraints in public housing allocation. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. 973–981.

[11] Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. 2011. The price of fairness. *Operations Research* 59, 1 (2011), 17–31.

[12] Omar Besbes, Yonatan Gur, and Assaf Zeevi. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*. 199–207.

[13] Miranda Bogen. 2019. All the Ways Hiring Algorithms Can Introduce Bias. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias. (2019).

[14] M. Bogen and A. Rieke. 2018. *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias.* Technical Report. Upturn.

[15] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia (Eds.). 2016. *Handbook of Computational Social Choice.* Cambridge University Press.

[16] James A Breaugh and Mary Starke. 2000. Research on employee recruitment: So many studies, so many remaining questions. *Journal of Management* (2000).

[17] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3992–4001.

[18] Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. 2014. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*. 379–387.

[19] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. *CoRR* abs/1810.08810 (2018). arXiv:1810.08810 http://arxiv.org/abs/1810.08810

[20] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[21] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[22] Pierre Desrochers. 2001. Local diversity, human creativity, and technological innovation. *Growth and Change* (2001), 369–394.

[23] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. 2013. Multi-armed bandit with budget constraint and variable costs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

[24] Jack Gillum and Ariana Tobin. 2019. Facebook Won't Let Employers, Landlords or Lenders Discriminate in Ads Anymore. (2019). https://www.propublica.org/article/facebook-ads-discrimination-settlement-housing-employment-credit

[25] Rachel Goodman. 2018. Why Amazon's Automated Hiring Tool Discriminated Against Women. (2018). https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against

[26] Moritz Hardt, Eric Price, , and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*. 3315–3323.

[27] Brent Hecht, Lauren Wilcox, Jeffrey P Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, and Bushra Anjum. 2018. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. *ACM Future of Computing Blog* (2018).

[28] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *CHI Conference on Human Factors in Computing Systems*. ACM, 600.

[29] Vivian Hunt, Dennis Layton, and Sara Prince. 2015. Diversity matters. *McKinsey & Company* (2015).

[30] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *International Conference on Machine Learning (ICML)*.

[31] Michael Kearns, Aaron Roth, and Saeed Sharifi-Malvajerdi. 2019. Average Individual Fairness: Algorithms, Generalization and Experiments. *arXiv preprint arXiv:1905.10607* (2019).

[32] Tallai Andherbertrobbins Lai and Herbert Robbins. 1985. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics* 6, 1 (1985), 4–22.

[33] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web (WWW)*. ACM, 661–670.

[34] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms.. In *WSDM*, Irwin King, Wolfgang Nejdl, and Hang Li (Eds.). ACM, 297–306. http://dblp.uni-trier.de/db/conf/wsdm/wsdm2011.html#LiCLW11

[35] Jérémie Mary, Romaric Gaudel, and Philippe Preux. 2015. Bandits and Recommender Systems. In *Machine Learning, Optimization, and Big Data*. 325–336.

[36] Kanishka Misra, Eric M Schwartz, and Jacob Abernethy. 2019. Dynamic Online Pricing with Incomplete Information Using Multiarmed Bandit Experiments. *Marketing Science* 38, 2 (2019), 226–252.

[37] Ritesh Noothigattu, Snehalkumar S Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D Procaccia. 2018. A voting-based system for ethical decision making. In *Proc. 32nd AAAI Conference on Artificial Intelligence*.

[38] Cathy O'Neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[39] Anshuka Rangi and Massimo Franceschetti. 2018. Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers' ability. In *International Conference on Autonomous Agents and MultiAgent Systems*. 1345–1352.

[40] J. Rawls. 1971. *A Theory of Justice.* Harvard University Press.

[41] F. Rossi and N. Mattei. 2019. Building Ethically Bounded AI. In *Proc. of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*.

[42] Cynthia Rudin. 2013. Predictive policing using machine learning to detect patterns of crime. *Wired Magazine* (2013).

[43] Candice Schumann, Samsara N. Counts, Jeffrey S. Foster, and John P. Dickerson. 2019. The Diverse Cohort Selection Problem. In *International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. 601–609.

[44] Candice Schumann, Zhi Lang, Jeffrey S. Foster, and John P. Dickerson. 2019. Making the Cut: A Bandit-based Approach to Tiered Interviewing. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[45] Candice Schumann, Zhi Lang, Nicholas Mattei, and John P. Dickerson. 2019. Group Fairness in Bandit Arm Selection. In *NeurIPS Workshop on Machine Learning and Causal Inference for Improved Decision Making*.

[46] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2219–2228.

[47] Keivan G Stassun, Susan Sturm, Kelly Holley-Bockelmann, Arnold Burger, David J Ernst, and Donna Webb. 2011. The Fisk-Vanderbilt Master's-to-PhD Bridge Program: Recognizing, enlisting, and cultivating unrealized or unrecognized potential in underrepresented minority students. *American Journal of Physics* 79, 4 (2011), 374–379.

[48] Richard S. Sutton and Andrew Barto. 2017. *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

[49] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[50] Wei Tang and Chien-Ju Ho. 2019. Bandit Learning with Biased Human Feedback. In *International Conference on Autonomous Agents and MultiAgent Systems*. 1324–1332.

[51] Pelin Vardarlier, Yalcin Vural, and Semra Birgun. 2014. Modelling of the Strategic Recruitment Process by Axiomatic Design Principles. In *Procedia–Social and Behavioral Sciences*. 374–383.

[52] Paul Voigt and Axel von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide* (1st ed.). Springer Publishing Company, Incorporated.

[53] Jingyan Wang and Nihar B Shah. 2019. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *International Conference on Autonomous Agents and MultiAgent Systems*. 864–872.

[54] Huasen Wu, Rayadurgam Srikant, Xin Liu, and Chong Jiang. 2015. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*. 433–441.

[55] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. 2016. Conservative bandits. In *International Conference on Machine Learning*. 1254–1262.