An Entreaty for Fairness Dynamics

Owain West[1]

[1]otwest@uwaterloo.ca

Abstract

Although most work on fairness in machine learning has heretofore focused primarily on single-shot classification or regression tasks, there is nascent literature on machine fairness dynamics, that is, the temporal evolution of fairness measures as data-driven algorithmic decisions and societal responses interact. We present a detailed survey of observational data from historical studies in the applied sciences, as well as theoretical fairness results from recent computer science, with the goal of synthesizing progress and insight from these two perspectives. We describe models which we believe comprise the current state-of-the-art among dynamical models of fairness, and discuss their efficacy as well as the motivations and assumptions which underpin their chosen fairness criteria. We argue that both the experimental and formal perspectives provide evidence for the necessity of dynamical considerations and the inadequacy of a single-shot framework in real-world fairness applications.

1 Introduction

The study of fairness in machine learning has received much recent academic and media attention, coinciding with a rising awareness that the data-based predictions made by machine learning systems can and do encode existing biases from the observed data into their decision rules. For example, [BCZ$^+$16] showed that word embeddings trained using the common word2vec model on the Google News dataset (see [MSC$^+$13]) represented significant gender stereotypes.

The formal study of fairness is difficult in that it exists at the turbulent junction of theory and practice; formal tools can only do so much in the face of biased data. It is the disparity between ideal notions of fairness and the biased nature of the observed world that necessitates this difficulty, rather than the fact that one is using a computer to make a decision. This is made formal in the next section.

Why Fairness is Hard

Numerous statistical definitions of fairness have been proposed in the machine learning literature, generally coinciding with accepted notions of fairness from the social science literature. Taking A to be a sensitive attribute, X a feature, Y a true label, and $\hat{Y}$ a predicted label[1], we have the following three fairness measures worth highlighting for their general acceptance:

Definition 1. ***Equal Opportunity*** *requires equal selection rates across sensitive attributes. Formally, this means that* $\mathbb{P}[\hat{Y} = 1 \mid A = 1] = \mathbb{P}[\hat{Y} = 1 \mid A = 0]$.

[1]Herein, all attributes, features, and labels are binary.

Definition 2. ***Demographic parity*** *requires equal true positive rates across sensitive attributes. Formally, this means that* $\mathbb{P}[\hat{Y} = 1 \mid A = 1 \wedge Y = 1] = \mathbb{P}[\hat{Y} = 1 \mid A = 0 \wedge Y = 1]$. *This is also called* balance for the positive class. *Similarly,* ***balance for the negative class*** *requires* $\mathbb{P}[\hat{Y} = 1 \mid A = 1 \wedge Y = 0] = \mathbb{P}[\hat{Y} = 1 \mid A = 0 \wedge Y = 0]$.

Definition 3. ***Calibration*** *by group requires that for all possible predicted scores* $y \in \hat{Y}$ *and classes* $a \in A$, $\mathbb{P}\{Y = 1 \mid \hat{Y} = y, A = a\} = y$. *When this score is used,* $Y, \hat{Y}$ *are often not binary.*

It is important to note that these are all *observational* criteria, in the sense that they are defined by probability distributions over the given features and labels (for more on observational criteria, see [BHN17]). This means that they necessarily cannot model causal factors directly.

Certain philosophical positions naturally give rise to these fairness definitions. [FSV16] introduces the mathematical notion of a *construct space*, which is meant to capture unobserved features which effect the observed world; decisions are based on the observed world[2]. They state two possible strong axioms: the WYSIWYG[3] axiom, that there is only an ϵ-bounded amount of *distortion* (intuitively, difference)[4] between the construct and observed spaces, and the WAE[5] axiom that assumes the philosophical position that groups are approximately equal in value in the construct space (the difference is again bounded by some $\epsilon > 0$). They show that under only the first axiom, calibration can always be achieved, whereas under the second axiom demographic parity can always be achieved.

These fairness criteria are not mutually compatible however, except in the most restricted settings. In [KMR16], Kleinberg et. al. show that it is only possible to concurrently satisfy calibration and positive/negative class balance if at least one of the following hold:

- Measurement is perfect: we know everyone's class label exactly. Fairness can be achieved by classifying everyone accurately.
- Equal base rates: both classes have the same percentage of their members in the true positive class. Fairness can be achieved by assigning everyone a uniform score equal to the base rate.

Of course, neither of these restrictions are reflected in the observed world.

Calibration by group is the criterion which is naturally optimized by decisions based on observational criteria (this is made formal in [LSH18]). As described in [LSH18], this implies that calibration by group should only be accepted as a fairness criterion if we are comfortable with the fairness of the underlying observational data. As the examples of Section 2 show, this is hardly a defensible position in many real-world settings.

Beyond the difficulty of concurrently satisfying contrasting notions of fairness, it may also not be possible to compose related fair decisions in a manner which is itself fair (see, eg, [DI18]). As the literature on fairness dynamics makes clear, temporal dynamics can result in unfair global effects under repeated application of a locally fair decision (see Section 3).

For a broad overview of fairness in machine learning, see [BHN17].

2 Empirical Lessons

The formal study of machine fairness dynamics is closely connected in its viewpoint to ample historical work in the social sciences literature which model and evaluate the dynamical effects

[2]We note that this distinction between extra-sensory and perceivable reality is an age-old one.
[3]What You See Is What You Get
[4]See [FSV16] for a formalism.
[5]We're All Equal

of organizational decision rules.

Redlining

Redlining was a practice overseen by the US government-sponsored Home Owners' Loan Corporation (HOLC) beginning in 1935 in which neighbourhoods were categorized by their "desirability", a distinction which was used to assess creditworthiness of the residents and then came to be reflected in home prices and social mobility among residents. The process primarily negatively affected Black Americans, as the assigned creditworthiness score was a simple proxy for race. This practice continued until the 1970s, when such open discrimination in lending was made illegal. In [AN16], Appel. et al consider the longterm effects of this practice. They show that relevant characteristics varied continuously across boundaries prior to the introduction of redlining maps, whereas by 1990 home prices within redlined neighbourhoods were approximately 5% lower than they would have been otherwise (controlling for factors such as home size, age, state of repair, sales demand, etc). They characterize this state as being the result of a dynamics wherein lower-scored neighbourhoods were less inhabited, leading to a feedback loop of increasing home vacancy and depreciating value. Their analysis, as well as other works (see eg [KDK06]), evidences that the fairness effects of a practice are coupled with community responses and therefore necessarily have associated temporal dynamics.

Affirmative Action

"Affirmative action" is a name given to a legal concept of fairness which corresponds to our definition of demographic parity. It was initially brought up in the American context as a method of increasing the representation of minorities in governmental positions, but has since had a wider application in industry hiring and academic admissions.

[KRB+85] is an early observational study of the efficacy of affirmative-action policies on the medical profession. The article notes that the number of individuals from disadvantaged groups who were practicing physicians had doubled between 1970 and 1985, but considers such a broad measure to be inadequate from the perspective of evaluating social good. Instead, the authors evaluate various more specific population-level effects, such as the distribution of primary-care specialists of each class and the percentage of members of each class who chose to treat patients from the same class. Looking at this broad and socially-justified swath of criteria, they claim that affirmative action had a positive effect on medical schools. It is important to note that the looseness of their analysis from the perspective of a statistician or computer scientist is not necessarily a failing: fairness is always with respect to issues of human interest, and therefore the evaluation of the efficacy of fairness measures may take such "soft" factors into account in a situation-dependent manner. This highlights the strong model-dependence of fairness considerations.

[CL93] is an early, more formally-principled study of the analytical efficacy of affirmative action policies. It gives an analytical model of hiring dynamics in which two populations, ex-ante equivalent, are observed by an employer. Its primary result is that it is possible for groups which are ex-ante equivalent to become caught in a negative feedback loop in which an initial imperfect observation has compound societal effects. If the employer initially observes one population to be less qualified than the other, the employer updates their beliefs about the populations, and adjusts hiring accordingly. The effect of this is modelled as a lowering of effort investment from the disadvantaged population, which in turn results in worse observation, and the process cycles in this way.

[KDK06] analyses the efficacy of a number of different programs which are meant to support the broader societal goal of the equalization of workplace demographics. They found that programs which require corporate responsibility (eg affirmative action programs and diversity committees) were most effective, followed by network-based programs (eg mentoring and networking programs), and with the least effective class of interventions being ones based on individual behavioral change (eg diversity training or feedback in performance evaluations). These findings highlight that the qualitative character of fairness interventions is relevant to long-term fairness dynamics.

3 Computational Models of Dynamics

In this section, we describe the modelling choices and results of recent scholarship on the fairness properties of formally-specified models. Because modelling choices are of utmost importance in evaluating fairness results, for clarity we delineate the discussion of each paper into individual sections discussing their modelling choices and analytical results. For readability and because of the high symbol-variance among the discussed papers, modelling choices are represented in English rather than in their mathematical formalism when doing so does not alter their content, and some symbolic choices are standardized across papers.

Reinforcement Learning in MDPs: [JJK$^+$16]

Modelling Choices

[JJK$^+$16] is the first study of fairness principles in a reinforcement learning context. It models the world as a discounted Markov decision process running for a determined number of steps $\langle S, A, P, R, T, \gamma \rangle$; here, S is the state-set, A the set of actions, $P : S \times A \to S$ and $R : S \to [0, 1]$ the transition and reward matrices, T the (possibly $\aleph_0$) number of steps, and γ the discount factor. A *snapshot* of an MDP is a single $(state, action, reward)$ triple. Actions are taken by an independent actor, who selects a *policy* π that maps the previously observed snapshots to a distribution over potential actions at every time t for $1 \leq t \leq T$. Each policy is associated with functions V^π, Q^π which take states (resp. actions) to their associated long-term discounted rewards; the rewards associated with the optimal policy are denoted V^*, Q^*.

The fairness definition used corresponds to *weak meritocracy*, in the sense that π is considered to be fair if it with high probability does not ever choose an action a' over an action a whenever a's long-term expected reward is higher than that of a'. They show that this weak-seeming fairness criterion can be overly restrictive, and present two weaker versions: the *approximate-choice fairness* criterion and the *approximate-action fairness* criterion. The former is a relaxation of the weak meritocratic fairness criteria which allows the policy to prefer a worse choice over a better option by a factor of α, whereas the latter insists that the learned policy prefer action a whenever the long-term reward of a is at least α more than that of a'.

Results

The primary results of [JJK$^+$16] pertain to the hardness of approximating fairness in the MDP model. To approximate the optimal solution to the MDP with discount factor γ[6], any

[6] [JJK$^+$16] calls notion of approximation being elided here ϵ-optimality. It corresponds to the idea that the learning algorithm should converge to a policy that visits states with reward arbitrarily close to the reward of the optimal states

approximate-action fair algorithm requires time exponential in $\frac{1}{1-\gamma}$. Morever, any algorithm satisfying weak meritocracy or approximate-choice fairness requires time exponential in the number of MDP states in order to do so.

Discussion

Given the described "weak" nature of the fairness criterion, it is perhaps surprising that the results are so negative. From one perspective, the fairness criterion is indeed weak: it allows, for example, two equivalently-rewarding actions to be selected with unequal probabilities, contradicting calibration. Although they are weak in this sense, the restriction that *no* decisions violating such fairness measures can be made in the pipeline seems to be itself a key limitation. Consider a two-class system in which both classes are equal in inherent value but (via some external process of discrimination) have serious observational disparity. Provided that one had outside information pertaining to the true equal-nature of the classes, it is hard to see why a more drastic intervention than allowed by the proposed fairness criteria would not be justified.

Single-Track Pipeline Fairness: [BKN$^+$17]

Modelling Choices

[BKN$^+$17] defines the notion of a n-stage *pipeline*. Formally, a pipeline on a set X is a $P_X(f,g) := \langle X, \{f_i\}_{1\leq i\leq T}, \{g_i\}_{1\leq j<T}\rangle$; the $f_i : X \times \prod_{j<i} D_j \to D_i$ and $g_i : D_i \to \{0,1\}$ are called the *decision* and *rule* functions[7]. At each stage, an evaluation is made according to the f_i, and based on that evaluation, a decision is made by g_i. The final decision is given by a function defined in terms of f_T if g_t returns 1 for all $1 \leq t < T$ and a special failure value $*$ otherwise. Intuitively, this corresponds to a decision-making pipeline in which an individual is scored at each stage by the decision functions f_i, and may progress to the next stage dependent on the evaluation of the rule-function g_i (itself a function of the value of f_i and all previous scores). Examples of pipelines which fit into this framework include those that filter participants (such as interview processes) and those that enact cumulative or history-dependent decisions (such as progression through the education system). Pipelines wherein the decision at a single timestep can depend on population-level decision statistics from previous timesteps are not compatible with this model.

Results

Under the strong assumption that a true final positive predicted class implies positive intermediary predicted classes along the pipeline, [BKN$^+$17] shows that sequential decisions multiplicatively preserve fairness ϵ-bounds on equal opportunity fairness. They mention that fairness may not be preserved when the aforementioned strong assumption is not valid.

Discussion

The central result is very weak, as it depends on there being no false negatives anywhere along the pipeline. It may be possible to control for such factors in a situation where a single organization is able to control the whole pipeline, by having high acceptance rates along the pipeline until the final decision. This is likely to be impractical.

[7]Take the empty product to be a singleton set, so that $X \times \prod \emptyset \cong X \times \{\emptyset\} \cong X$

Local Dynamical Effect of Fair ML: [LDR$^+$18]

Modelling Choices

[LDR$^+$18] is a seminal paper in machine learning fairness. It takes the perspective that it is the downstream effect of fairness interventions, rather than the selection rate of a fairness-aware model, which is of primary interest. In order to model this, groups A, B are given score distributions π_A, π_B over a finite set $\mathcal{X}$; these distributions represents the groups' inherent characteristics. An external institution selects *policies* $\tau_A, \tau_B : \mathcal{X} \to [0, 1]$, which correspond to selection rates by group and score. The model assumes that the institution can calculate the utility of its own policies, and moreover that the average effect of a policy on the members of a group $j \in \{A, B\}$ with some initial score x is also calculable. Say that a policy results in *long-term improvement* (resp. decline) for a group if the change in average score for that group increases (resp. decreases) as the result of a policy. Say the policy results in *stagnation* if there is no such change. Say that a policy causes *active harm* if it causes long-term decline in a population relative to the base rate, *relative harm* if it causes decline relative to the change caused by the institution's unconstrained optimal policy, and *relative improvement* in the case where it causes more improvement than does the unconstrained optimal policy.

Results

The results of [LDR$^+$18] are manifold. Foremost, they show that both equal opportunity and demographic parity can result in any one of decline, stagnation, or increase. Under some conditions – eg that the two feature distributions π_A, π_B are equal up to translation – more can be shown: demographic parity can cause active harm (by overallocating to the disadvantaged class) while equal opportunity cannot. Moreover, demographic parity is always more "generous" to the disadvantaged class than equal opportunity, and so much so that it is never less generous than the unconstrained classifier. Under the assumption that the institution has more to lose than the individual does in ratio, the authors additionally show that it is not possible for the institution's unconstrained policy to cause active harm. They formally show that threshold policies are preferable from a fairness perspective in this model. They introduce the *outcome curve*, a graph of selection rate against demographic improvement, and show that it must be concave in this model. This visual tool makes it easy to interpret their main theorems; for a visual, we refer the reader to [LDR$^+$18].

Discussion

The framework presented in [LDR$^+$18] is more general than previously published models, and has the benefit of directly addressing the effects of varying fairness criteria. It is highly interpretable, as outcome curves can be plotted for various fairness criteria, and qualitatively evaluated against the institutional utility plotted against the same selection rate. This sort of analysis allows policy-makers, value-ethicists, and institutional leaders to directly estimate the relative fairness and institutional effects of varying fairness measures. Such analysis is a necessary precondition for the concurrent moral and fiscal evaluation of a classifier.

The model in [LDR$^+$18] has the property of only dealing with local fairness dynamics – the changes in outcome are measured against a single application of the decision policy to the population. The authors argue that this is necessary for most applications of fairness dynamics, as longer-term effects may be overridden by external societal factors. The extent to which this is true could be a matter of debate.

There are a number of (perhaps necessary) downsides to this approach. Although the single-decision effects are more easily integrated into a fairness plan for an individual decision

maker, the focus on single-turn effects can tell a different story than a longer-term view of dynamics might. This is made clear in [DSA+20].

A few assumptions inherent in the model are worth mentioning. The first is that the average effect on members of a class who have a given initial score is assumed to be a known deterministic function. Similarly, the model assumes that the institution is able to deterministically model its own policy utility. One can imagine numerous scenarios when such deterministic information is not available. Lastly, many results (including the non-harm of unconstrained optimization) depend on the assumption that the institution takes more risk than the individual does, ie that it has relatively higher loss associated with a false positive than the individual has for the expected change in score associated with a false negative.

Simulating Global Dynamics: [DSA+20]

Modelling Choices

[DSA+20] puts the focus directly on the challenge of accurately modelling fairness dynamics. It introduces `ml-fairness-gym`, a Python package implementing OpenAi's `gym` reinforcement learning environment (see [BCP+16]). `ml-fairness-gym` as described in [DSA+20] contains three *environments*, corresponding to various toy models of fairness. One of these models is that of [LDR+18]; the others detailed in [DSA+20] are an attention-allocation model (ie predictive policing, similar to [EFN+17]), and a college admissions environment meant to evaluate affirmative action effects. These environments and the standardized framework that supports them allow for simulation under varying policies and parameter values, and provide a means of evaluating the longer-term dynamical effects of fairness interventions in various models by allowing counterfactuals to be tested.

Results

The results of this paper, especially the simulation results relevant to the [LDR+18] paper, are of particular interest. First, [DSA+20] shows that the qualitative character of a fairness effect can be different in the long-term than after a single step as in [LDR+18].

In particular, simulations in a lending environment like in [LDR+18] show that while equal-opportunity can do relative harm by overlending (resulting in a lower long-term credit score for individuals of a disadvantaged class), such an effect may not necessarily be detrimental from the perspective of a disadvantaged class itself. Indeed, the disadvantaged group *still receives the same number of loans* in the long-term case *even though* their credit scores have been systematically lowered. From the perspective of the "disadvantaged group", it is not clear whether this really is a disadvantage at all.

Secondly, the assumptions of [LDR+18] may not hold up exactly in the real-world. We noted the assumption that the institutional lender in [LDR+18] is assumed to take more risk than the applicant for most of their analysis. This holds in most real-world cases, but may not in edge cases. In particular, individuals with a maximum credit score of 800 take more risk than the bank does in applying for a loan[8], and such effects ultimately result in the permanent lowering of credit scores to below the maximum threshold.

Thirdly, optimizing for local fairness dynamics may not be optimal for global fairness dynamics. In particular, implementing local equal opportunity constraints at each instance does not result in an optimal summed global true positive rate. This is shown both empirically in the simulation of the [LDR+18] model, and analytically as a formal theorem.

[8]From the sole perspective of credit-score maximization, of course.

Discussion

The simulations described in [DSA+20] give clear evidence that fairness considerations are only properly understood in the dynamical context, and that this longer-term context introduces additional complexities. It highlights the importance of simulating toy models of fairness-relevant situations in order to gain insight into the disparate effects that varying measures can have, and cautions that even models that consider short-term dynamics may not adequately capture the limiting effects of a given policy. There has since been one more paper published, [ASHS19], contributing a new environment to `ml-fairness-gym`. More work of this sort is warranted.

Other Models

[LWH+20] analyzes a model in which individuals invest in some preparatory good (eg education) in a rational manner dependent on their expected utility gain, and then an institution makes a decision (eg hiring) based on the qualification level presented by an individual. This is very similar to the setting of [CL93] in the historical literature on affirmative action in hiring. [MOS18] is another recent paper which considers a highly similar situation, studying the effects of affirmative-action type policies on a two-stage selection process.

[HC17] considers a dynamical model of hiring fairness in which the labour market is split into two sub-markets: a temporary labour market in which the disadvantaged class is systematically boosted in order to attain some fairness goal, and then a permanent labour market into which workers transition after some time. They give recursive solutions characterizing the distributions of good workers in both the temporary and permanent labour markets of their model.

[EFN+17] considers a model of predictive policing in which there are two types of crime: reported crime (which is a function of neighbourhood only) and discovered crime (which is a function of police allocation). By modelling this using a generalized Polya urn model, they transfer standard results about the steady-state of such urns into a fairness context. In particular, they show that this model captures many of the (unfair) features of the `PredPol` policing algorithm.

[ZKTL19] discusses the effects of a decision rule on differentially-distributed classes in a group setting where negative decisions influence individuals to remove themselves from the group. This is similar to the filtering pipeline context of [BKN+17]. They assume that the decision rule has a retention effect which is monotone in the sense that if group A makes up a larger (resp. smaller) fraction of the population at some time t as compared to the fraction represented by A at time t', then the retention rates after the corresponding one-shot decisions will be larger (resp. smaller). Under the retention monotonicity assumption, Zhang et. al. show that a group's proportion can, unsurprisingly, only change monotonically in the general case: once a population's representation begins to grow (resp. shrink) it continues to do so. They also show that the retention monotonicity assumption is satisfied in the case where the retention rate for groups is a decreasing function of some function of the group's observational decision parameters. One example of a model which fits this description relates to software design: the retention rate of users is a decreasing function of the error rate of a piece of software.

There is a growing body of work that applies similar techniques in the study of the multi-armed-bandit problem (a special case of an MDP with only a single state) in a reinforcement learning context. For research in this direction, see [JKMR16, BBC+16, BBK17, KKM+17, LRD+17, KMR+18]. For a formal treatment of the multi-armed bandit setting see [Sli19].

4 Discussion

Fairness is not important in a vacuum, and there is no utility at all in working to make machine-learning systems (or any decision-making system) work independently of whether one has tied their shoelaces or not. Rather, sensitive attributes are meaningfully sensitive because they correspond to groupings which have been historically disadvantaged, and thus for which historical decisions – which are embedded in observational data – do not reflect the (idealized) equal nature of the sensitive and nonsensitive groups. This highlights that the utility of fairness as an applied concept[9] is a function of its *effects* on the relevant population rather than a pure function of its direct selection rate across populations. This perspective is largely elided by single-shot fair classification, which obviates the fundamental difficulty: we live in a reactive world, rife with feedback loops and extant observational bias.

The results of the described papers make the following general points clear:

- Measures that promote static fairness may not promote dynamical fairness ([DSA+20]).
- Dynamical fairness has difficulties analogous to those faced by static fairness: there are similar formal impossibility results, although approximations are possible ([KRZ18, AKRZ20, DSA+20]).
- Dynamical fairness also suffers from the added difficulty in modelling, as social responses to fairness measures are at best weakly understood ([KDK06]) and small changes in model parameters can have large dynamical effects ([DSA+20]).
- Local dynamical fairness effects can be quantified under reasonable assumptions ([LDR+18]), but long-term fairness effects may be seriously different even under the same assumptions ([DSA+20]).

Taken together, these provide further evidence for the inadequacy of considering fairness measures at a single timestep: static treatment of fairness ignores these necessary difficulties, rather than facing them head-on.

5 Conclusion

To the extent that all that matters for fairness is the uniform application of rules which are understood to be fair (and thus necessarily uniform for observationally equivalent individuals of given sensitive class[10]), it is wholly irrelevant whether the ultimate arbiter of the decision be human or machine, provided the same decision is made[11]. The study of the realizability of fairness goals is thus independent from the study of machine fairness interventions, and the fact that the former has been studied extensively in the social science literature on the basis of observational criteria therefore can and should inform current work. The historical literature in social sciences is rife with examples of the central importance of dynamical considerations in fairness applications, and moreover provides numerous case-studies in which dynamical considerations cause myopic fairness goals to compose in potentially-surprising ways. Recent analytical results in computer science have characterized some of these nonobvious dynamical effects, and simulations can serve as a useful tool for experimentation or qualitative analysis in the absence of formal guidelines. Together, historical effects, formal theorems, and simulated

[9] Which, again, we strongly believe to be its only utility.

[10] This condition is equivalent to the specification that decisions must be made according to potentially class-specific thresholds, whether probabilistic or discrete, in order to have a chance at being fair.

[11] Of course, this is not the case when considering an explainability or accountability perspective; this solely refers to the observationally-dependent nature of the decision.

experiments all highlight the importance of careful modelling of the feedback effects of fairness constraints.

References

[AKRZ20] Eshwar Ram Arunachaleswaran, Sampath Kannan, Aaron Roth, and Juba Ziani, *Pipeline interventions*, ArXiv **abs/2002.06592** (2020).

[AN16] Ian Appel and Jordan Nickerson, *Pockets of poverty: The long-term effects of redlining*, SSRN **4** (2016).

[ASHS19] James Atwood, Hansa Srinivasan, Yoni Halpern, and D Sculley, *Fair treatment allocations in social networks*, 2019.

[BBC+16] Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach, *Exploring or exploiting? social and ethical implications of autonomous experimentation in ai*, Workshop on Fairness, Accountability, and Transparency in Machine Learning (2016).

[BBK17] Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi, *Mostly exploration-free algorithms for contextual bandits*, ArXiv **abs/1704.09011** (2017).

[BCP+16] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba, *Openai gym*, 2016.

[BCZ+16] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*, CoRR **abs/1607.06520** (2016).

[BHN17] Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness in machine learning*, NIPS Tutorial (2017).

[BKN+17] Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian, *Fair pipelines*, CoRR **abs/1707.00391** (2017).

[CL93] Stephen Coate and Glenn C Loury, *Will affirmative-action policies eliminate negative stereotypes?*, The American Economic Review **83** (1993), 1220–1240.

[DI18] Cynthia Dwork and Christina Ilvento, *Fairness under composition*, CoRR **abs/1806.06122** (2018).

[DSA+20] Alexander D'amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern, *Fairness is not static: deeper understanding of long term fairness via simulation studies*, FAT'20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency **abs/1910.04123** (2020), 525–534.

[EFN+17] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian, *Runaway feedback loops in predictive policing*, Conference on Fairness, Accountability, and Transparency, Proceedings of Machine Learning Research (2017), 1–12.

[FSV16] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian, *On the (im)possibility of fairness*, CoRR **abs/1609.07236** (2016).

[HC17] Lily Hu and Yiling Chen, *A short-term intervention for long-term fairness in the labor market*, CoRR **abs/1712.00064** (2017).

[JJK+16] Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, and Aaron Roth, *Fair learning in markovian environments*, CoRR **abs/1611.03071** (2016).

[JKMR16] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth, *Fairness in learning: Classic and contextual bandits*, CoRR **abs/1605.07139** (2016).

[KDK06] Alexandra Kalev, Frank Dobbin, and Erin Kelly, *Best practices or best guesses? assessing the efficacy of corporate affirmative action and diversity policies*, American sociological review **71** (2006), 589–617.

[KKM+17] Sampath Kannan, Michael J. Kearns, Jamie Morgenstern, Mallesh M. Pai, Aaron Roth, Rakesh V. Vohra, and Zhiwei Steven Wu, *Fairness incentives for myopic agents*, CoRR **abs/1705.02321** (2017).

[KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, *Inherent trade-offs in the fair determination of risk scores*, CoRR **abs/1609.05807** (2016).

[KMR+18] Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu, *A smoothed analysis of the greedy algorithm for the linear contextual bandit problem*, CoRR **abs/1801.03423** (2018).

[KRB+85] Stephen Keith, M Robert, August Bell, Albert Swanson, and null Williams, *Effects of affirmative action in medical schools: a study of the class of 1975*, New England Journal of Medicine **313** (1985), 1519–1525.

[KRZ18] Sampath Kannan, Aaron Roth, and Juba Ziani, *Downstream effects of affirmative action*, CoRR **abs/1808.09004** (2018).

[LDR+18] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt, *Delayed impact of fair machine learning*, CoRR **abs/1803.04383** (2018).

[LRD+17] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes, *Calibrated fairness in bandits*, CoRR **abs/1707.01875** (2017).

[LSH18] Lydia T. Liu, Max Simchowitz, and Moritz Hardt, *Group calibration is a byproduct of unconstrained learning*, CoRR **abs/1808.10013** (2018).

[LWH+20] Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes, *The disparate equilibria of algorithmic decision making when individuals invest rationally*, FAT'20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency **abs/1910.04123** (2020), 381–391.

[MOS18] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro, *From fair decision making to social equality*, CoRR **abs/1812.02952** (2018).

[MSC+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean, *Distributed representations of words and phrases and their compositionality*, CoRR **abs/1310.4546** (2013).

[Sli19] Aleksandrs Slivkins, *Introduction to multi-armed bandits*, CoRR **abs/1904.07272** (2019).

[ZKTL19] Xueru Zhang, Mohammad Mahdi Khalili, Cem Tekin, and Mingyan Liu, *Long term impact of fair machine learning in sequential decision making: representation disparity and group retention*, CoRR **abs/1905.00569** (2019).