Group Fairness in Bandits with Biased Feedback

Candice Schumann  
University of Maryland  
College Park, USA  
candiceschumann@gmail.com

Nicholas Mattei  
Tulane University  
New Orleans, USA  
nsmattei@gmail.com

Zhi Lang  
University of Maryland  
College Park, USA  
willylang1996@gmail.com

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

ABSTRACT
We propose a novel formulation of group fairness with biased feedback in the contextual multi-armed bandit (CMAB) setting. In the CMAB setting, a sequential decision maker must, at each time step, choose an arm to pull from a finite set of arms after observing some context for each of the potential arm pulls. In our model, arms are partitioned into two or more sensitive groups based on some protected feature(s) (e.g., age, race, or socio-economic status). Initial rewards received from pulling an arm may be distorted due to some unknown societal or measurement bias. We assume that in reality these groups are equal despite the biased feedback received by the agent. To alleviate this, we learn a societal bias term which can be used to both find the source of bias and to potentially fix the problem outside of the algorithm. We provide a novel algorithm that can accommodate this notion of fairness for an arbitrary number of groups, and provide a theoretical bound on the regret for our algorithm. We validate our algorithm using synthetic data and two real-world datasets for intervention settings wherein we want to allocate resources fairly across groups.

KEYWORDS
Group fairness; fair bandits; contextual bandits; human collaboration

ACM Reference Format:

Knowing that one may be subject to bias is one thing; being able to correct it is another.

Jon Elster

1 INTRODUCTION
In many online settings, a computational or human agent must sequentially select an item from a slate, receive feedback on that selection, and then use that feedback to learn how to select the best items in the following rounds. Within computer science, economics, and operations research circles, this is typically modeled as a multi-armed bandit (MAB) problem [46]. Examples include algorithms for selecting what advertisements to display to users on a webpage [28], systems for dynamic pricing [29], and content recommendation services [44]. Indeed, such decision-making systems continue to expand in scope, making ever more important decisions in our lives such as setting bail [31], making hiring decisions [6, 37], and policing [35]. Thus, the study of the properties of these algorithms is of paramount importance as highlighted by Chouldechova and Roth [11] motivating priorities for fairness research in machine learning.

In the basic MAB setting, there are n arms, each associated with a fixed but unknown reward probability distribution [2, 23]. 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. The goal of the agent is to maximize the total collected reward over time. A generalization of MAB is the contextual multi-armed bandit (CMAB) where the agent observes a $d$-dimensional context along with the observed rewards to choose a new arm. In the CMAB problem, the agent learns the relationship between contexts and rewards and selects the best arm [1].

Yet, the use of MAB- and CMAB-based systems often results in behavior that is societally repugnant. Sweeney [47] noted that queries for public records on Google resulted in different 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. Following that initial report similar instances continue to be observed, both in the bandit setting and in the general machine learning world [31]. In lockstep, the academic community has begun developing approaches to tackling issues of (un)fairness in learning settings. We have an opportunity to identify and understand why the data we have may be causing the bias.

A Computing Community Consortium (CCC) report on fairness in ML identified that most studies of fairness are focused on classification problems [11]. These works define a statistical notion of fairness, typically a notion of equal treatment of equals [33], and propose algorithms to abide by these constraints. Two issues identified by Chouldechova and Roth [11] that we address in this paper are extensions to notions of group fairness and looking at fairness in online dynamic systems, e.g., CMABs. We address these gaps by formalizing and providing an algorithm for fairness with biased feedback when the arms of the bandit can be partitioned into groups. Direct applications of our work including scenarios...
discussed within the AAMAS community like aiding the allocation of human resources in talent sourcing [38].

The recent AI100 study [26], whose goal is to take a broad and long-term look at the opportunities and pitfalls for AI researchers, has highlighted the need to develop systems that work with humans, providing oversight, transparency, and explanation. Our bandit formulation is one step towards creating more human-centered AI [41], a new area of study that seeks to understand and balance computer automation and autonomy with the level of human control in a given system. Many of the negative applications of MAB based systems we have discussed so far too often occur because there is too much autonomy given to the system, and it optimizes away from what humans or society considers desirable. By explicitly modeling the underlying bias term, we hope to improve computer aided decision making by understanding and mitigating the dangers that can occur when there are excessive levels of human control or excessive levels of computer autonomy; leading to systems that are more transparent, auditable, and trustworthy.

**Running Example.** As a running example throughout the paper, imagine the position of an agent at a bank or a lender on a micro-lending site. Here, the agent must sequentially pick loans to fund. In many cases, such as the micro-lending site Kiva, a user is presented with a slate of potential loans they may fund when they log in and this slate is generated by a recommender system [45]. Each of these loans, i.e. arms, has a context which includes attributes of the applicant (e.g., personal statement, repayment history, business plan). The loans can also be partitioned into sets of $m$ sensitive attributes, e.g. location, race, or gender. In the simplest case, assume we have two female applicants and two male applicants on the slate at a given time. We also assume that when pulling an arm from, for example, a female applicant, there is some societal bias introduced into the reward. Yet, in many settings (and, as we assume in this work), the average true (i.e., unbiased) reward across groups is equal. We want to balance the number of times the agent selects women versus men given this societal bias built into the feedback.

While we use loans as our running example, our notion of regret could be extended to a number of other areas including recent work in MAB problems on hiring situations [39], including the recent AAMAS Blue Sky Paper by Schumann et al. [38] specifically calling for the community to contribute to fair hiring. One could imagine a situation where hiring decisions are made w.r.t. a short-term reward signal that is biased,

\[1\] versus a longer-term reward of performance which is less biased, e.g., via an end-of-year review that is based on a more quantitative metric such as on-the-job performance. A similar argument can be made about school admissions or matching workers to online tasks in a crowdwork setting.

**Contributions.** We propose a novel formulation of group fairness in the contextual multi-armed bandit (CMAB) setting. In our model, arms are partitioned into two or more sensitive groups based on some protected feature, e.g., race. Despite the fact that there may be differences in expected payout between the groups, we may wish to ensure some form of fairness between picking arms from the various groups. Our goal is to capture the phenomena where we want to balance the arms being pulled from both groups and (learn to) ignore societal bias generated by sensitive group membership. We define two novel notions of reward and regret to capture implicit societal bias: proportional parity and equal group parity. We provide a novel algorithm that can accommodate these notions of fairness for an arbitrary number of groups, learn the societal bias term itself, and provide bounds on the regret for our algorithm. We validate our algorithms using synthetic data and real-world datasets for intervention settings wherein we want to allocate resources fairly across protected groups.

## 2 RELATED WORK

Fairness in machine learning has become one of the most active topics in computer science [11]. The idea of using formal notions of fairness, i.e. axioms or properties, to design decision schemes has a long history in economics and political economy [33, 53]. Typically within ML research, fairness is operationalized using the Rawlsian idea that similar individuals should be treated similarly; formally extended to the classification setting by Dwork et al. [15], who provided algorithms to ensure individual fairness at the cost of the utility of the overall system. Their work underscores that in many cases statistical parity is not sufficient to ensure individual fairness, as we may treat groups fairly but in doing so may be very unfair to some specific individual. Determining when, how, and if to define fairness is an ongoing discussion with roots well before the time of computer science [44]; indeed, it is known that many natural conditions for fairness cannot be achieved in tandem [21]. Still, group fairness is found in many fielded systems [4, 51], and we focus on it here.

The study of fairness in MAB was initiated by Joseph et al. [19], who showed for both MAB and CMAB one can implement a fairness definition where within a given pool of applicants, e.g., college admission or mortgages, a worse applicant is not favored over a better one, despite a learning algorithm’s uncertainty over the true payoffs. However, Joseph et al. [19] only focus on individual fairness, and do not formally treat the idea of group fairness. Individual fairness, in some sense, group fairness taken to an extreme, where every arm is its own singleton group; it offers strong guarantees, but under strong assumptions [5, 20].

Celis et al. [9] propose a bandit-based approach to personalization where arm pulls are constrained to fit some probability distribution defined by a fairness metric such as demographic parity. For example, when recommending news articles, their algorithm provides personalized articles from both left and right sources. Their formulation is perhaps closest in the literature to our formulation as they deal with group fairness, however it does not explicitly assume biased feedback. Instead, it enforces a fair probability distribution without learning about the bias present in the data.

There are a number of other recent studies of fairness in the MAB literature. Chen et al. [10] investigate a task allocation setting with a fairness constraint that captures a minimum rate at which

---

1Class-based bias presents itself within seconds of an in-person interview; see https://news.yale.edu/2019/10/21/yale-study-shows-class-bias-hiring-based-few-seconds-speech.

2A full version of this paper, complete with an appendix containing proofs and additional experiments, can be found at https://arxiv.org/abs/1912.03802. We will also periodically update this work should typos or other errors be found; if you see any, please feel free to reach out! Code to reproduce the experiments is also available at https://github.com/candieschumann/groupfairoptimization.
a task is assigned to a particular arm; their model is quite general and captures the adversarial and some non-stationary settings. Liu et al. [27] look at fairness between arms under the assumption that arm reward distributions are similar (another interpretation of equal treatment of equals). Patil et al. [32] define fairness such that each arm must be pulled for a predetermined required fraction over the total available rounds. Claure et al. [12] use the MAB framework to distribute resources amongst teammates in human-robot interaction settings; again, fairness is defined as a pre-configured minimum rate that each arm must be pulled. Hussain et al. [17] take a more theory-oriented approach to a similar setting, proposing a multi-agent variant of a stochastic MAB setting with a Nash social welfare definition of fairness.

Since preliminary versions of this work were presented [38, 40] there have been several papers that have investigated similar problems. Wang et al. [50] look at fairness of exposure in CMAB base systems, specifically focusing on similarity of merit, which is more in line with individual rather than the group fairness we consider here. Tang et al. [49] consider a setting inspired by liver transplantation where the objective is to trade off a more egalitarian, max-min, policy in allocating opportunities for surgeons to gain experience in liver transplant training. Finally, Ron et al. [34] investigate a setting of allocating opportunities to sub-populations in a corporate decision making setting where each arm needs to pulled at least a budgeted number of times, but where the cost of allocating an opportunity to a non-optimal arm is known in advance. Interestingly, their algorithm also achieves a $T^{2/3}$ regret, similar to our results.

One needs to be careful when appealing to purely statistical metrics for ensuring fairness. As argued by Corbett-Davies and Goel [13], simply setting our sights on a form of classification parity, i.e., metrics for ensuring fairness. As argued by Corbett-Davies and Goel [13], simply setting our sights on a form of classification parity, i.e., ensuring individual fairness, $\hat{\beta}_i$ a vector of coefficients; if arm $i$ is pulled at timestep $t$, the following reward is returned: $r_{i,t} = \hat{\beta}_i \cdot x_{i,t} + \mathbb{1}[i \in P_1] \psi_{p_1} \cdot x_{i,t} + N(0, 1)$, where $\mathbb{1}[i \in P_1] = 1$ when $i \in P_1$ and 0 otherwise, and $\psi_{p_1}$ is a societal or systematic bias against group $P_1$. Note that $\psi_{p_1}$ is a vector for the non-sensitive group. Hence, the underlying biased utility function can be written as $f(x_{i,t}) = \hat{\beta}_i \cdot x_{i,t} + \mathbb{1}[i \in P_1] \psi_{p_1} \cdot x_{i,t}$.

Using our running example, let’s assume that the down payment reward received has some bias against the male applicants compared to the female applicants, while the final repayment does not. Note that the final repayment is not measured after accepting a loan and is only measured much later. The loan agency should then take the bias into account while learning what ‘good’ applications look like. Or, in a hiring setting, an applicant may have a biased interview (initial reward) while their true performance is measured only after working for a year (later true reward).
We define true regret for pulling an arm $a$ at time $T$ as

$$R^*(T) = \sum_{t=1}^{T} f^*(x^t_{i,t}) - f^*(x_{a,t})$$  \hspace{1cm} (3)$$

where $i^*$ is the optimal arm to pull at time $t$ and $f^*(x_{i,t})$ is the true reward with no bias terms $\psi_{P_i} \cdot x_{i,t}$. We also assume that the average true reward (with no bias) for group $P_i$ should be the same as the average reward for group $P_j$ in Equation 1, which would return the regret on the biased reward function $f(x_{i,t})$. In the loan agency example, this real regret $f^*(x_{i,t})$ would measure the regret of the final repayments instead of the biased down payment regret.

One can view the societal bias term $\psi_i$ that we learn for some group $i$ as our algorithm learning how to automatically identify and adjust for anti-discrimination for group $i$ compared to all other groups. Anti-discrimination is the practice of identifying a relevant feature in data and adjusting it to provide fairness under that measure [13]. One example of this, discussed by Dwork et al. [15], Joseph et al. [19], and in the official White House algorithmic decision making statement [30], comes up in college admissions. Given other factors, specifically income level, some colleges weight SAT scores less in wealthy populations due to the presence of tutors while increasing the weight of working-class populations [3]. While in these admissions settings the adjustments may be ad-hoc, we learn our bias term from data. Past work has compared the vector $\beta_i$ learned for each arm as akin to adjusting for these biases [15]. While this is true at an individual level, our explicit modeling of bias allows us to discover these adjustments at a group level.

4. GROUP FAIR CONTEXTUAL BANDITS

In this section, given our new definition of reward (Equation 2) and correspondingly new definition of regret (Equation 3), we present the algorithm GroupFairTopInterval (Algorithm 1) which takes societal bias into account. We also give a bound on its regret in this new reward and regret setting. Subsequently, we briefly describe the algorithm.

In GroupFairTopInterval, each round $t$ is randomly chosen with probability $\frac{1}{T}$ to be an exploration round. The exploration round randomly chooses an arm to pull.

The remaining rounds become exploitation rounds, where linear estimates are used to pull arms. GroupFairTopInterval learns two different types of standard OLS linear estimators [22]. The first is a coefficient vector $\hat{\beta}_{i,t}$ for each arm $i$ (line 7). Additionally, GroupFairTopInterval learns a group coefficient vector $\hat{\psi}_{P_i,t}$ for each group $P_j$ (lines 4 and 5). To calculate these coefficient vectors, the algorithm keeps track of previous arm pull rewards for each arm $i$ at every timestep $t$ in a vector $Y_{i,t}$, and the corresponding contexts for each arm pull in a matrix $X_{i,t}$. A similar vector $\mathcal{Y}_{P_i,t}$ and matrix $\mathcal{X}_{P_i,t}$ is kept for both groups $P_j$. As mentioned previously, we treat $P_i$ as the sensitive group of arms. An arm $i$ in the non-sensitive group $P_j$ has a reward estimation of $\hat{\beta}_{i,t} \cdot x_{i,t}$, while an arm $i$ in the sensitive group $P_i$ has a bias corrected reward estimation of $\hat{\beta}_{i,t} \cdot x_{i,t} - \hat{\psi}_{P_i,t} + \hat{\psi}_{P_i,t}$.

For each arm $i$, the algorithm calculates confidence intervals $w_{i,t}$ around the linear estimates $\hat{\beta}_{i,t} \cdot x_{i,t}$ using a Quantile function $Q$ (line 9). This means that the true utility (including some bias) falls within $[\hat{\beta}_{i,t} \cdot x_{i,t} - w_{i,t}, \hat{\beta}_{i,t} \cdot x_{i,t} + w_{i,t}]$ with probability $1 - \delta$ at every arm $i$ and every timestep $t$. Similarly, for each group $P_i$ and context $w_{i,t}$ for a given arm $i$ at timestep $t$, the algorithm calculates a confidence interval $b_{P_i,t}$ using a Quantile function $Q$ (lines 4 and 5). This means that the true group utility (or true average group utility) falls within $[\hat{\psi}_{P_i,t} \cdot x_{i,t} - b_{P_i,t}, \hat{\psi}_{P_i,t} \cdot x_{i,t} + b_{P_i,t}]$ with probability $[1 - \delta]$. Using the confidence intervals $w_{i,t}$ and $b_{P_i,t}$, and the linear estimates $\hat{\beta}_{i,t} \cdot x_{i,t}$ and $\hat{\psi}_{P_i,t} \cdot x_{i,t}$, we calculate the upper bound of the estimated reward for each arm $i$ (lines 15 and 17), pulling the arm with the highest upper bound (line 18).

Algorithm 1 GroupFairTopInterval

Require: $\delta$, $P_j$, $P_i$
1: for $t = 1 \ldots T$ do
2:  With probability $\frac{1}{T}$, play $i_t \in \{1, \ldots, n\}$ and observe reward $y_{i_t,t}$
3:  otherwise:
4:    $\hat{\psi}_{P_i,t} := \left( X_{P_i,t}^T X_{P_i,t} \right)^{-1} X_{P_i,t}^T Y_{P_i,t}$
5:    $\hat{\psi}_{P_i,t} := \left( X_{P_i,t}^T X_{P_i,t} \right)^{-1} X_{P_i,t}^T Y_{P_i,t}$
6:  for $i = 1 \ldots n$ do
7:    $\hat{\beta}_{i,t} := \left(X_{i,t}^T X_{i,t}\right)^{-1} X_{i,t}^T Y_{i,t}$
8:    $F_{i,t} := \mathcal{N}(0, \sigma^2 X_{i,t} \left( X_{i,t}^T X_{i,t} \right)^{-1} X_{i,t}^T)$
9:    $w_{i,t} := Q_{F_{i,t}} \left( \frac{\delta}{\sqrt{n}} \right)$
10:   if $i \in P_i$ then
11:       $T_{P_i,t} := \mathcal{N}(0, \sigma^2 X_{i,t} \left( X_{P_i,t}^T X_{P_i,t} \right) X_{i,t}^T)$
12:       $T_{P_i,t} := \mathcal{N}(0, \sigma^2 X_{i,t} \left( X_{P_i,t}^T X_{P_i,t} \right) X_{i,t}^T)$
13:       $b_{P_i,t} := Q_{T_{P_i,t}} \left( \frac{\delta}{2 \sqrt{n}} \right)$
14:       $b_{P_i,t} := Q_{T_{P_i,t}} \left( \frac{\delta}{2 \sqrt{n}} \right)$
15:       $\hat{u}_{i,t} := \hat{\beta}_{i,t} \cdot x_{i,t} + w_{i,t} - \hat{\psi}_{P_i,t} \cdot x_{i,t} + b_{P_i,t}$
16:       $\hat{u}_{i,t} := \hat{\beta}_{i,t} \cdot x_{i,t} + w_{i,t}$
17:       $u_{i,t} := \hat{\beta}_{i,t} \cdot x_{i,t} + w_{i,t}$
18:       Play $\text{argmax}_i u_{i,t}$ and observe reward $y_{i_t,t}$

Returning to our running example, using GroupFairTopInterval, the loan agency would learn a down payment reward function for each of the arms, i.e., a coefficient vector $\beta_i$ where $i \in \{\text{young female arm}, \text{young male arm}, \text{older female arm}, \text{older male arm}\}$, as well as the group average coefficients for the gender-grouped arms, $\hat{\psi}_{P_i}$, for male and female. Using the gender-grouped coefficients, expected rewards for male arms are reweighted to account for the bias in down payment.

Standard algorithms like TopInterval4 would choose an arm $i = \text{argmax} \hat{\beta} \cdot x_{i,t} + w_{i,t}$, ignoring societal bias (Equation 2, leading to a larger true regret (Equation 3)). Note that GroupFairTopInterval can be extended to multiple groups by defining an overall average reward.

4 A variant of the contextual bandit LinUCB by Li et al. [24]
**GroupFairTopInterval** is fair—in the context of our group fairness definitions—and satisfies the following theorem. Appendix B of the full paper provides a detailed, complete proof.

**Theorem 1.** For two groups $P_1$ and $P_2$, where $P_1$ has a bias offset in rewards, **GroupFairTopInterval** has regret

$$R^*(T) = O \left( \sqrt{\frac{d \ln T}{\nu}} T^{2/3} + \left( \frac{d}{\delta^2} \left( \ln^2 \frac{2 \nu T}{\delta^2} + \ln d \right) \right)^{2/3} \right).$$

Proof Sketch. We start by proving two lemmas. The first of which states that with probability at least $1 - \delta$:

$$\left| \hat{\beta}_{i,t} \cdot x_{i,t} - (\beta_i \cdot x_{i,t} + 1) \left[ i \in P_1 \right] \varphi_{P_1} \cdot x_{i,t} \right| \leq w_{i,t}$$  \hspace{1cm} (4)

holds for any $i$ at time $t$. Similarly, the second states that with probability at least $1 - \delta$:

$$\left| \hat{\beta}_{i,t} \cdot x_{i,t} - \beta_i \cdot x_{i,t} \right| \leq w_{i,t}$$ \hspace{1cm} (5)

holds for any group $P_j$, any arm $i$, and at any timestep $t$. By combining these two lemmas, we can see that arms should be treated fairly.

The regret for **GroupFairTopInterval** can be broken down into three terms:

$$R^*(T) = \sum_{t: t \text{ is an exploit round}} \text{regret}(t)$$  \hspace{1cm} (6)

+ \sum_{t: t \text{ is an exploit round and } t < T_1} \text{regret}(t)

+ \sum_{t: t \text{ is an explor round and } t \geq T_1} \text{regret}(t).$$  \hspace{1cm} (7)

First, for any $t$ we have:

$$\sum_{t < T_1} \frac{1}{T_1^{1/3}} = \Theta(t^{2/3}).$$  \hspace{1cm} (8)

We then show that the number of rounds $T_1$ after which we have sufficient samples such that the estimators are well concentrated is:

$$T_1 = \Theta \left( \min_a \left( \frac{d n \lambda_{\min,\nu}}{\lambda_{\min,\nu}} \left( \ln^2 \frac{2 \nu T}{\lambda_{\min,\nu}} + \ln d \right) \right)^{1/2} \right).$$  \hspace{1cm} (9)

Finally, we bound the third term in Equation 6 as follows:

$$\sum_{t: t \text{ is an explor round and } t \geq T_1} \text{regret}(t)$$  \hspace{1cm} (10)

$$\leq O \left( \sqrt{\frac{d n \ln \frac{2 \nu T}{\lambda_{\min,\nu}}}{\lambda_{\min,\nu}} T^{2/3} + \delta T} \right).$$  \hspace{1cm} (11)

Combining Equations 6, 8, 9, and 10, we have Theorem 1. \qed

Note that we can extend Algorithm 1 to $m$ groups. In this setting, we make the strong assumption that true rewards are centered about $\rho$ defined by the user.\(^5\) In this adaption of the algorithm, we set the upper bound radius for arm $i$ as:

$$\hat{u}_{i,t} = \hat{\beta}_{i,t} \cdot x_{i,t} + w_{i,t} + \rho - \hat{\varphi}_{P_i} \cdot x_{i,t} + \lambda_{\min,\nu}.$$  \hspace{1cm} (12)

where $i \in P_j$. We then have the following theorem for multiple groups:

\(^5See Appendix B.2 for further details.

**Theorem 2.** For $m$ groups $P_1, \ldots, P_m$, **GroupFairTopInterval** (Multiples Groups) has regret

$$R^*(T) = O \left( \sqrt{\frac{d n \ln \frac{2 \nu T}{\lambda_{\min,\nu}}}{\lambda_{\min,\nu}} T^{2/3} + \left( \frac{d}{\delta^2} \left( \ln^2 \frac{2 \nu T}{\delta^2} + \ln d \right) \right)^{2/3} \right).$$

where $l = \min \lambda_{\min,\nu}$, with $\lambda_{\min,\nu}$ the smallest eigenvalue of $X_{i,t}X_{i,t}^T$; and $L > \max \lambda_{\max} (X_{i,t}X_{i,t}^T)$.

5 EXPERIMENTS

To empirically evaluate **GroupFairTopInterval**, we perform experiments on synthetic data to demonstrate the effects of various parameters, and on real datasets to demonstrate how **GroupFairTopInterval** performs in the wild. In each of these sections we compare to **TopInterval**, due to Li et al. [24], **NaiveFair** (See Section 5.1), and **IntervalChaining**, due to Joseph et al. [19].

5.1 **NaiveFair**

One popular definition of group fairness in classification is the notion of demographic parity. Formally, given a protected demographic group $A$, we want:

$$\Pr(\hat{Y} = 1 | A = 0) = \Pr(\hat{Y} = 1 | A = 1),$$  \hspace{1cm} (12)

where the probability of assigning a classification label $\hat{Y} = 1$ does not change based on the sensitive attribute class $A$. Demographic parity is important when ground truth classes $Y$ are extremely noisy for sensitive groups due to some societal or measurement bias. Assume that we have a classifier that predicts whether an individual should receive a loan where our sensitive attribute $A$ is binary gender. Demographic parity states that the probability of getting a loan should be the same for males ($A = 0$) and females ($A = 1$).

In converting this definition of demographic parity to the multi-armed bandit setting, we alter the definition to be that the probability of pulling an arm $a$ does not change based on group membership $P_j$:

$$Pr(\text{pull } a | a \in P_j) = Pr(\text{pull } a | a \in P_k).$$  \hspace{1cm} (13)

Continuing our running example, assume we are a loan agency. The loan agency receives $4$ applications at every timestep $t$: an applicant from a young female, an applicant from a young male, an applicant from a older female, an applicant from an older male; we must choose one application to grant at each timestep. After granting a loan the loan agency receives a down payment on that loan as reward. This reward is then used to update the estimates of whether or not a “good” loan application was received for the pulled arm. Assume that the loan agency wants to act fairly using the binary sensitive attribute of gender. Then, the probability that the loan agency chooses a female applicant at timestep $t$ should be the same as the probability of choosing a male applicant.

A naive algorithm to enforce this definition of fairness is defined in Algorithm 2. We first pick from the groups uniformly at random, and then apply a regular CMAB algorithm like **TopInterval**\(^6\) or

\(^6**TopInterval** is a variant of LinUCB by Auer et al. [2].
would randomly pick between male or female, and then choose the within the group. Using our running example, NaiveGroupFair

Algorithm 2 NaiveGroupFair

Require: δ, P₁, P₂
1: for t = 1 . . . T do
2: P ← Randomly choose group P₁ or P₂.
3: Pull arm in P based on TopInterval

ContextualThompsonSampling [1] to choose which arm to pull within the group. Using our running example, NaiveGroupFair would randomly pick between male or female, and then choose the best applicant between the younger and older pair.

5.2 Synthetic Experiments

In each synthetic experiment, we generate true coefficient vectors βᵢ by choosing coefficients uniformly at random for each arm i. Contexts at each timestep t are chosen randomly for each arm i. Bias coefficients ϕ₁ are set uniformly at random with mean μ = 10. Seeds are set at the beginning of each experiment to keep arms consistent between algorithms.

We run four different types of experiments: 7) a) Varying the total budget for pulling arms (T) while setting number of arms n = 10, error mean μ = 10, number of sensitive arms equal to 5, and context dimension d = 2 (Figures 2a and 1a). b) Varying the total number of arms n while setting total budget T = 1000, error mean μ = 10, ratio of sensitive arms to 50%, and context dimension d = 2 (Figures 2b and 1b). c) Varying the error mean μ while setting total budget T = 1000, number of arms n = 10, number of sensitive arms equal to 5, and context dimension d = 5 (Figures 2c and 1c). d) Varying the number of sensitive arms while setting total budget T = 1000, number of arms n = 10, error mean μ = 10, and context dimension d = 2 (Figures 2d and 1d).

The plots in Figure 1 show the percentage of times an algorithm pulled a sensitive arm over the full budget T. In order to be fair, the percentage of sensitive arms pulled should be proportional to the number of sensitive arms, i.e., when there are 5 sensitive arms out of 10 total, the percentage of sensitive arms pulled is roughly 50%. Figure 2 shows the perceived regret that includes bias ϕ as solid lines, and real regret that corrects bias (see Equations 2 and 3) as dashed lines. Algorithms with low real regret are considered ‘good’.

Figure 1a shows that once exploration is over, GroupFairTopInterval pulls sensitive arms roughly 50% of the time, matching the 50% of sensitive arms. Figure 2a shows that GroupFairTopInterval performs comparably on real regret as TopInterval performs on biased regret. This means GroupFairTopInterval should be used over TopInterval in contexts where bias is anticipated. NaiveFair performs poorly in the context of societal bias.

Figure 1b illustrates that IntervalChaining becomes more group fair as the number of arms increase. This is because many arms are chained together and therefore, arms are chosen uniformly at random. Figure 2b illustrates this random picking of arms as real regret and biased regret increases dramatically for IntervalChaining.

As expected, Figure 1c illustrates that when the error mean μ is large, both IntervalChaining and TopInterval choose fewer sensitive arms. This leads to a high real regret as shown in Figure 2c. Following Kleinberg et al. [21], Figure 2c also suggests that one cannot have both individual and group fairness in a scenario with high mean error. The randomness in NaiveFair leads to a very high regret for both perceived regret and real regret.

Figure 1d demonstrates the fairness property of proportionality. The percentage of sensitive arms pulled by GroupFairTopInterval matches the number of sensitive arms. As shown in Figure 2d,

![Figure 1: Percentage of total arm pulls that were pulled using sensitive arms.](image1)

![Figure 2: Regret for synthetic experiments. The solid lines are regret given the rewards received from pulling the arms (including the group bias). The dashed lines is the true regret (without the group bias).](image2)

![Algorithm 2 NaiveGroupFair](algorithm2)
the number of sensitive arms does not affect the real regret of GroupFairTopInterval.

5.3 Experiments on Real-World Data

After exploring GroupFairTopInterval on synthetic data, we move on to using both the Philippines family income and expenditure dataset on Kaggle and the ProPublica COMPAS dataset. When one looks at the gender and age breakdown in the family income dataset, one can see that quite often female heads of households make more money than males in the Philippines. This is most likely due to the large number of Filipino women who work out of the country; it is estimated that up to 20% of the GDP of the Philippines is actually remittances from these overseas—primarily female—workers. In fact, almost 60% of overseas workers are women and 75% of these women are between the ages of 25 and 44. In the COMPAS dataset ProPublica observed a societal bias over recidivism risk scores for African-Americans.

**Experimental Setup.** Given the skew of high income coming from female head of households in the family income dataset, we treat the binary ‘Household Head Sex’ feature as the sensitive attribute. To create arms, we split up households based on ‘Household Head Age’ bucketed into the following five groups: (8, 27], (27, 45], (45, 63], (63, 81], and (81, 99]. We then have 10 different arms (for example, two arms would be Female head of household between 8 and 27, and Male head of household between 8 and 27).

Similarly, we treat African-American individuals from the COMPAS dataset as the sensitive attribute. We create arms by splitting up households based on the three age categories found in the data. We therefore have six different arms.

At each timestep $t$, we randomly select an individual from each arm. The context vector is the remaining features where any nominal features are transformed into integers. After an arm is pulled, a reward of the household income (for the family income dataset) or violent decile score (for COMPAS) is returned. We use these datasets for illustrative purposes.

Results. We see the same behavior of arm pulls in the real world data. Figures 3a and 3c show that after a period of exploration, the percentage of sensitive arms (male-grouped arms) pulled gets very close to 50%, matching the proportion of sensitive-grouped arms.

Figures 3b and 3d are perhaps more interesting. Since we cannot measure the “real” regret without the bias we assumed from the sensitive-grouped arms, we consider the gap between GroupFair-TopInterval and TopInterval as the price of fairness. The gap in regret is small compared to the increase in percentage of sensitive arms pulled. However, the gap in regret for NaiveFair is large in comparison. This suggests that explicitly learning a societal bias term will help in biased settings with low price to perceived regret. Note that there is a difference between regret scales for the two different datasets. This is due to the family income and expenditure dataset reporting regret in income, while the COMPAS dataset reports regret in recidivism score.

6 GENERAL DISCUSSION & ETHICAL IMPLICATIONS OF THE WORK

This work was directly motivated by research into bias found in machine learning models. There have also been calls to action for more research to be done on bias mitigation in online learning settings, specifically in multi-armed bandit settings and related areas such as recommender systems. Directly addressing these calls, in this work we propose a method of alleviating societal or measurement bias introduced into reward feedback. Using our CMAB model should help mitigate biased behaviors found in bandit systems currently in use.

Additionally, as noted by O’Neil [31], models can provide a biased feedback loop. We hope that by incorporating a societal bias term we can learn something about the bias that is being introduced. The coefficient vector will show which features are incorporating bias into the model. This allows users to address these features outside of the model and potentially find the sources of the societal bias. We do note that addressing societal bias and fixing the solution is a nontrivial task, the societal bias term provides the initial step of measurement.

On the other hand, as noted by Schumann et al. [38], Shneiderman [42], and many others, humans should still be active participants in decision making. If models such as our CMAB model are used to replace more and more human decision makers, this could...
have unintended and potentially negative medium- and long-term side effects. All models should be monitored for biased feedback loops in the given contexts that they are being used [31].

Choosing a particular definition of fairness—conditioned on deciding that it is even appropriate to formally define a notion of fairness in the first place—is a morally-laden decision. We note that, as machine learning practitioners, in many societally-relevant applications it is paramount that we maintain an open dialogue with stakeholders. In this work, we analyze a sequential decision-making system under one particular definition, group fairness; it is certainly not the case that this is a one-size-fits-all solution that would be deployable without receiving input from that larger set of stakeholders. Indeed, recent research [36] shows that non-expert users may have vastly varying degrees of comprehension of different definitions of fairness, and that the degree of comprehension may be a function in part of education level and other features that may correlate with measures of marginalization; this hints that the consequences of incorporating fairness definitions into machine-learning-based systems may not be uniformly understood by participants, and indeed that those participants who may be impacted the most by that change could comprehend that potential impact the least. In an allocative system like the one we describe in the main paper, nuanced considerations must be considered.

Our new definitions of reward (Equation 2) and regret (Equation 3) for the MAB setting provide an opportunity to look at biased data in a new light. In many cases, ground truths provided during learning are noisy with respect to sensitive groups. Additionally, debiased ground truths may be very expensive to receive or may take a long time to acquire. For instance, if looking at loans, true rewards of repayment may take years to receive. Or, for example, in hiring—the true reward of hiring an individual may take over a year to estimate, while the initial estimate may be influenced by a hiring team’s unconscious bias over features such as ethnicity, gender, or orientation.

Our proposed algorithm, GroupFairTopInterval, learns societal bias in the data while still being able to differentiate between individual arms. Previous solutions relied on setting ad-hoc thresholds, requiring some form of quota, or choosing groups uniformly at random. While it is true that GroupFairTopInterval can easily be extended to a case where we know that the average of a group is a constant offset from the other group. That being said, defining such offsets raises a host of other ethical questions. For instance, in the US, the EEOC (Equal Employment Opportunity Commission) posposed that the ratio of the most favored group compared to the other group is certainly not the case that this is a one-size-fits-all solution that would be deployable without receiving input from that larger set of stakeholders. Indeed, recent research [36] shows that non-expert users may have vastly varying degrees of comprehension of different definitions of fairness, and that the degree of comprehension may be a function in part of education level and other features that may correlate with measures of marginalization; this hints that the consequences of incorporating fairness definitions into machine-learning-based systems may not be uniformly understood by participants, and indeed that those participants who may be impacted the most by that change could comprehend that potential impact the least. In an allocative system like the one we describe in the main paper, nuanced considerations must be considered.

Our new definitions of reward (Equation 2) and regret (Equation 3) for the MAB setting provide an opportunity to look at biased data in a new light. In many cases, ground truths provided during learning are noisy with respect to sensitive groups. Additionally, debiased ground truths may be very expensive to receive or may take a long time to acquire. For instance, if looking at loans, true rewards of repayment may take years to receive. Or, for example, in hiring—the true reward of hiring an individual may take over a year to estimate, while the initial estimate may be influenced by a hiring team’s unconscious bias over features such as ethnicity, gender, or orientation.

Future work could expand GroupFairTopInterval to enforce individual fairness within groups. Intersectional group fairness is also important to look at in the MAB setting where more than one type of sensitive attribute needs to be protected. Additionally, other group fairness definitions such as Equalized Opportunity should be converted to the MAB setting [16]. Another interesting direction for future work is to mix ideas from the study of budget constrained bandits [14, 52] with our fairness definitions. We have also assumed individual arms have fixed group membership; generalizing to a setting where memberships in protected groups may change at every timestep $t$ would fit more real world applications.

ACKNOWLEDGMENTS

Dickerson and Schumann were supported by NSF CAREER Award IIS-1846237, NSF Award CCF-1852352, NSF Award SMA-2039862, NIST MSE Award #20126334, DARPA GARD #HR00112020007, DARPA SL3-CMD #S4761, DoD WHS Award #HQ003420FO0035, and ARPA-E DIFFERENTIATE Award #1257037. Mattei was supported by NSF Awards IIS-RI-2007955 and IIS-III-2107505. We thank Aviva Prins and Aravind Srinivasan for helpful comments on earlier versions of this paper, and the anonymous reviewers for helpful comments and constructive critiques.

REFERENCES

Theoretical Computer Science Conference. ACM, 214–226.


