Policy Fairness and Unknown Bias Dynamics in Sequential Allocations

Meirav Segal
University of Oslo
Oslo, Norway
meiravs@ifi.uio.no

Anne-Marie George
University of Oslo
Oslo, Norway
annemage@ifi.uio.no

Christos Dimitrakakis
University of Neuchatel
Neuchatel, Switzerland
christos.dimitrakakis@unine.ch

ABSTRACT

This work considers a dynamic decision making framework for allocating opportunities over time to advantaged and disadvantaged individuals, focusing on the example of college admissions. Here, individuals in the disadvantaged group are assumed to experience a societal bias that limits their success probability. Bias dynamics dictate how the societal bias changes based on the current allocation of opportunities. We model this environment as a Markov Decision Process (MDP) and empirically examine the purely utility maximising policy in terms of fairness. We demonstrate the influence of the bias dynamics on long-term fairness of allocations, and analyse the interplay between utility and policy-fairness for different dynamics under different optimisation parameters. We consider the cases of known and unknown bias dynamics. For known dynamics, we show that a short horizon view presents fairness as a trade-off for utility, but a long horizon view reveals that the two are aligned. Moreover, we suggest that when the dynamics are unknown, the approach towards epistemic uncertainty may also affect fairness, and should be considered when designing fair decision making models.

CCS CONCEPTS

• **Computing methodologies** → **Markov decision processes**; **Sequential decision making**.

KEYWORDS

Long term fairness, Fairness-utility trade-off, Policy fairness

ACM Reference Format:
Meirav Segal, Anne-Marie George, and Christos Dimitrakakis. 2023. Policy Fairness and Unknown Bias Dynamics in Sequential Allocations. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23), October 30–November 01, 2023, Boston, MA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3617694.3623262

1 INTRODUCTION

AI models for allocation problems have become prevalent in many applications, such as lending [10], hiring [37] and policing [36]. These problems are characterised by a decision maker (DM) allocating limited resources among a population in order to maximise some objective. As these applications are considered high risk systems [17], recent efforts to provide fair allocations incorporate fairness constraints, e.g., guaranteeing a proportional portion of resources to protected groups [14]. While most of these endeavors focus on static settings (single allocation), the community has recently shifted towards considering sequential settings (sequence of allocations over time [54, 68]). We provide a group fairness definition for sequences of allocations and use it to experimentally analyse the fairness of utility-maximising allocations along the example of college admissions.

Sequential processes are particularly interesting to study because they may introduce *feedback effects*, so that current decisions may change the future population distribution. We analyse the college admissions example under such a feedback effect. In these dynamic processes, well-intended attempts to increase fairness might lead to negative long-term effects for the disadvantaged or protected group [34, 44]. Hence, there is a need to measure fairness in new ways which take into account the dynamics of the process. This has been studied largely under the assumption of full knowledge of the underlying dynamics [9]. Yet, typically dynamics are unknown and must be learned by interaction with the environment.

Although past works on sequential decision making address the problem of learning dynamics [19], some only use a point estimate of the dynamics to make decisions and do not take uncertainty into account. However, if the DM maintains a probabilistic belief over possible dynamics, it is possible to obtain better policies [11]. This view also allows the DM to be risk sensitive with respect to epistemic uncertainty. This approach towards uncertainty over the dynamics has, to the best of our knowledge, not been analysed under fairness measures. As DMs in real life applications might tend to favour cautious decisions, we consider policies maximising both expected utility and an approximate lower bound. We empirically analyse how these approaches towards uncertainty impact fairness.

We use college admissions as a running example to motivate our modeling choices and the relevance of our analysis. [1] Higher education is a key step for many career paths. As such, access to higher education is crucial for self fulfillment and financial security. Unfortunately, there are still sub-populations with reduced opportunities due to societal biases: discouraging environments, lack of role models and internalised stereotypes could lead to reduced chances of success through insufficient skill development or self-handicapping [26, 62]. Hence, DMs may try to equalise group participation through *affirmative action*, such as setting a lower acceptance bar for disadvantaged groups. For example, in the Norwegian admission system, some study programs provide bonus

EAAMO '23, October 30–November 01, 2023, Boston, MA, USA

ACM ISBN 979-8-4007-0381-2/23/10.
https://doi.org/10.1145/3617694.3623262

[1] Note that our framework is also applicable to other domains such as hiring and lending.

points based on gender (e.g. for men in nursing and for women in engineering) [52]. In practice, this translates to a lower acceptance threshold for that gender. Such actions could potentially generate more role models and provide investment incentive, which might encourage further skill development. However, depending on the societal dynamics, they might also have negative effects. For example, lowering the bar entails the admission of less qualified group members with reduced chances of graduation, leading to a lower success rate within this group. This might increase the bias by reinforcing stereotypes. In this sense, affirmative actions may or may not be aligned with *preferential treatment*, i.e., setting different thresholds to increase the success chances of the disadvantaged group.

College admissions are thus a fitting example for a sequential decision process with feedback in which fairness plays a key role while the effects of our decisions on future populations are unknown. We model this setting as a Markov Decision Process (MDP) and experimentally find *admission policies* that allocate study opportunities for any given state of the applicant population. Here the DM is assumed to be interested in maximising the number of successful students.

Our main contributions are as follows:

(1) **Affirmative Actions and Preferential Treatment:** In section 5 we show that for college admissions, different societal bias dynamics influence when and how preferential treatment is employed by a utility-maximising DM. Also, for some dynamics preferential treatment is aligned with affirmative actions while for others it is not. This emphasises the importance of understanding the dynamics before enforcing measures such as affirmative actions.
(2) **Utility-Fairness Trade-Off:** We measure the fairness of a policy by a weighted sum of state fairness over a fixed time horizon (see section 3). For known dynamics, we show in section 6.1 that when considering both short and long horizons, placing more weight on future rewards by increasing the discount factor leads to policies that are fairer. However, DMs considering a short horizon would see this as a trade-off for utility, while for DMs considering a long horizon the utility is aligned with fairness.
(3) **Epistemic Uncertainty and Fairness:** Lastly, in section 6.2 we show that when the dynamics (transition functions) are unknown, under a fixed (low) discount factor, a comparison between uncertainty-aware policies shows that a policy which performs well in terms of utility might perform worse in terms of fairness.

2 RELATED WORK

While some contributions have been made to analyse fairness in sequential decision making settings without feedback (e.g. [56, 57]), we focus on such settings with feedback, i.e., where the population is influenced by the decisions that were taken in the past. In such settings, many papers have recently explored the long-term effect of policies with and without static (per step) fairness constraints on the fairness towards sub-populations [2, 13, 32, 42, 48, 66]. Others also considered constraints that take into account the entire sequence of actions [19, 64]. In contrast, we do not impose fairness as a constraint, but merely analyse the fairness of utility maximising policies. As we do, some also consider the use of MDPs as a method for modeling the sequential process [9, 58, 61]. In our setting, the states of the MDP represent the societal bias and the transition function represents its dynamics.

Unknown Dynamics. Previous works mostly assume knowledge of the dynamics, with several exceptions learning the dynamics from historical data or through interactions with the environment [8, 50, 59, 64]. These works, however, do not address the *uncertainty* while estimating the transition function. One example [64] first explores using a uniform policy, then estimates the transition function according to the observations and finally optimise the policy with respect to this estimate. In contrast, we take a Bayesian approach and consider several samples of likely transition functions. A more recent example [67] uses online RL to find a policy, for initially unknown dynamics, which maximizes some objective under fairness constraints. They conclude that long-term fairness constraints drive the system towards policies that sacrifice short-term reward for a better utility-fairness trade-off in the long-term. In this paper, we show that such policies can be reached without fairness constraints, only by adapting the horizon of the decision maker.

Risk Under Uncertainty. Under epistemic uncertainty (uncertainty regarding the underlying MDP model), it is common to assume a Bayesian approach, marginalise over possible model parameters and then maximise expected utility. Risk-sensitivity can be added by using a non-linear utility function for returns. However, alternative notions of risk aversion have been used, also in Markov decision processes [65]. In reinforcement learning, recent progress in distributional methods [5] has led to the development of new methods for managing aleatory risk [7]. Such methods may take into account the distribution of the return [40, 41, 60]. We, however, focus on epistemic uncertainty, which Bayesian methods [21] suit rather well. Epistemic uncertainty has been studied in this context for risk-averse DMs [16, 51], improved utility and robustness [11].

Fairness Under Uncertainty. The use of uncertainty-aware policies brings about concerns regarding fairness under such methods. Epistemic uncertainty has been addressed so far in a Bayesian [12] and a PAC setting [31]. For unknown population distributions, Wen et al. address the uncertainty with respect to their estimations and find that risk-sensitive policies could have different long-term fairness implications [64]. Heidari et al. analyse both group discrimination and individual fairness under a risk-aversion model, where the uncertainty is due to the concept of a veil of ignorance [23] and Nokhiz et al. consider the concept of precarity and risk aversion from the individual's perspective [43]. Weber et al. learn the delayed impact of actions as unknown rewards based on a static dataset with the goal of enforcing long-term fairness [63]. We, on the other hand, consider uncertainty of the dynamics from the DM's perspective and its fairness implications at the population level.

Model Choices. Our setting is similar to that of Heidari and Kleinberg [24], who formalize the process of intergenerational mobility between groups with different socioeconomic status. In their model, individuals may shift groups based on their performance in opportunities granted to them. For instance, individuals who are admitted to university and graduate successfully, have increased chances of

earning higher salaries and their children to be more advantaged. A similar setting was proposed by Acharya et al., who model the college admission process for an advantaged and disadvantaged group [1]. The two groups have the same talent distribution (as in [24]) but they differ in their wealth. Here, the DM only observes a signal combining both the talent and the wealth, and decisions affect the future wealth distributions of the two groups. Under this setting, the authors analyse the effects of several interventions. In our setting, the groups are fixed, but our decisions can have an impact at the population level, by changing societal biases. A model for college admissions was proposed by Mouzannar et al., which also models the feedback with respect to populations [42]. They consider dynamics based on selection rates while we also analyse dynamics based on success rates. A similar analysis was performed in a model simulating the labour market, where the long-term fairness was evaluated for different kinds of interventions [53]. In our model, the feedback is induced directly by the decisions and we assume full knowledge of the population features. This is in contrast to work assuming the feedback is (partly or in full) due to unbalanced exploration (also known as the selective labels problem in supervised learning [4]) [13, 15, 36]. Moreover, we do not evaluate specific interventions, but observe the actions generated by utility-maximising policies.

Other Approaches. A different approach towards long-term fairness comes from causality literature, where the fairness of the dynamic process can be defined in terms of changes to a causal graph over time [27]. Furthermore, Creager et al. use a causal modeling framework for estimating outcomes of fairness intervention under unknown dynamics [8]. The phenomena of feedback loops in automated decision making has been addressed in the setting of supervised learning under the concept of Performative Prediction [46]. Several papers have addressed fairness concerns under this framework [39, 45, 47], yet we take a reinforcement learning approach. Other fairness consideration in the reinforcement learning framework were also suggested, such as for multi-armed bandits [33] and contextual bandits [30]. The latter, as well as an extension to MDPs [28], studies fairness as meritocracy: never probabilistically favour an action with a lower long-term reward over an action with a higher long-term reward. A notion of state-visitation fairness for MDPs is proposed in [20], where each state of the MDP should be visited with a pre-specified minimum frequency.

3 PRELIMINARIES

In this paper, we analyse the effect of different model parameters on the utility maximising policy, its utility and fairness, with policy optimisation performed using standard algorithms. Specifically, we would first like to test our hypothesis, that for our dynamic system, there is no trade-off between fairness and utility, only a trade-off between short-term and long-term views. Later, we will analyse the utility-fairness trade-off under different policy choices in the presence of uncertainty. To this end, we first define the DM's utility and then define a fairness measure for a policy. The following definitions fit a MDP setting where we have states S, actions A, rewards R and a transition function τ.

3.1 Utility

We define the utility of the DM as a discounted sum of rewards achieved by acting according to a (deterministic) policy π from a given start state s_0:

$$U^\pi(s_0) = \sum_{t=0}^{T} \gamma^t R(s_t, \pi(s_t)),$$

where $\gamma \in [0, 1)$ is the discount factor which determines the weight we give to future rewards. The closer γ is to 1, the more weight we place on future rewards. When the start state is unknown, we integrate over the possible start states s_0 through some start state distribution c. That is, we compute $U^\pi = \int_S U^\pi(s_0)\,\mathrm{d}c(s)$.

3.2 Fairness as Consequentialism

In static settings, fairness is usually measured with respect to immediate outcomes, e.g., measuring demographic parity of a given allocation. In dynamic settings, we can also examine fairness with respect to the resulting population state s_{t+1}. For example, one can measure the difference between the feature distributions of sub-populations in the new state [58]. To capture both outcome types, we can define the *immediate fairness outcome* $f_\tau(s_t, a_t)$ as some function of a_t and s_t under the transition function τ.

To evaluate the fairness of a policy, a possible approach would be to consider the fairness of a state reached after a fixed time T (with the possibility of $T = \infty$), as was studied in the past (e.g. [32, 35, 42, 55]). Yet, using this measure we cannot differentiate between policies along the entire time period. For example, we cannot differentiate two policies that reach an unbiased state after 20 steps, while one discriminates in every step but the last, and the other discriminates in the first step but is unbiased afterwards.

Instead, we define the fairness of a policy as the weighted sum of state fairness scores, according to the expected path induced by the policy. Formally, let us assume we have a function $f_\tau : S \times A \to \mathbb{R}$ such that for every state of the world and every action, we get a real number indicating the immediate fairness outcome. We define the fairness of a policy π according to weights $\{w_i\}_{i=0}^T$ for the time steps, visited states $\{s_i\}_{i=0}^T$ and their actions $\{a_i\}_{i=0}^T$ according to transition function τ starting from start state s_0 as:

$$F^\pi(s_0) = \mathbb{E}_\pi[\sum_{t=0}^{T} w_t f_\tau(s_t, a_t)]. \tag{1}$$

When the start state is unknown, we integrate over the possible start states to measure the fairness of a policy. That is, we compute:

$$F^\pi = \int_S \mathbb{E}_\pi[\sum_{t=0}^{T} w_t f_\tau(s_t, a_t)]\,\mathrm{d}c(s) \tag{2}$$

This is a generalisation of previously suggested notions where weights and f-function are fixed to specific values, such as demographic parity and equal opportunity for MDPs [64], or discounted sum of fairness costs in recommender systems [19]. Note that, if a discount factor γ is used as a weight, the fairness of a deterministic policy in a deterministic environment takes the same shape, $F^\pi(s_0) = \sum_{t=0}^{T} \gamma^t f_\tau(s_t, a_t)$, as the utility function where fairness scores correspond to rewards.

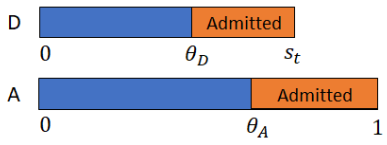


Figure 1: The success distribution and admission thresholds for the disadvantaged (D) and advantaged (A) group.

This fairness measure thus allows to identify the fairest policy in the same way as a utility-maximising policy. When replacing rewards by the state-action fairness measure f, we can readily apply any algorithm that finds an optimal policy to find a fairness maximising policy.

4 USE CASE: COLLEGE ADMISSIONS

We model a MDP in which the DM repeatedly allocates college admission slots to a fixed proportion ($\alpha \in [0,1]$) of the applicants. The DM's reward per round is the fraction of successful students among the entire population and the DM's utility is the discounted sum of total rewards over time.

We consider a population of applicants partitioned into two disjoint groups: disadvantaged (D) and advantaged (A). We assume that these groups represent a constant affiliation (such as race) so there are no transitions between groups. More specifically, we assume that group fractions ($\phi_D, \phi_A \in [0,1]$ with $\phi_D + \phi_A = 1$) remain constant over time. The fraction of the population that is disadvantaged, ϕ_D, is hence a fixed parameter of the MDP. In addition, we assume that group affiliation has no influence over the innate ability of individuals, which is uniformly distributed for both groups ($q_i \sim U[0,1]$ for individual $i \in A \cup D$). Formally, our MDP is a tuple $\langle S, A, R, \tau \rangle$ with state space S, action space A, reward function $R : S \times A \to \mathbb{R}$ and transition function (dynamics) $\tau : S \times A \to S$.

States. The state space is $S = [0,1]$. The state $s_t \in S$ is the societal bias factor at time t. In our model, societal bias influences the skill development of the disadvantaged group members, leading to lower chances of success: The success probability for an advantaged group member is equal to their innate ability $p_A(i) = q_i$, while for a disadvantaged group member, the success probability is their ability multiplied by the current bias factor $p_D(j) = s_t q_j$. In effect, the current state is translated to the upper bound of the success distribution of the disadvantaged group, i.e., $p_D \sim U[0, s_t]$ as visualised in Figure 1. We assume the MDP to be fully observable, i.e., the DM can observe the success probability distribution of candidates from either group at each time t.

Actions. Our action space is $A = [0,1]$, such that action a_t at time t is the admission threshold θ_D for the disadvantaged group. Note that because the fraction of admitted students is fixed to α, the admission threshold θ_A for the advantaged group follows from θ_D. More specifically, $\theta_A = \frac{(1-\alpha)s_t - a_t\phi_D}{s_t\phi_A}$ as explained in more detail in the appendix. A candidate is admitted if their success probability is over the group's threshold. Having different thresholds allows for *affirmative action* policies, e.g., see Figure 1 where the threshold of the disadvantaged group θ_D is lower than that of the advantaged group θ_A.

Rewards. The action a_t and state s_t determine the fraction of admitted and successful students from each group at time t, i.e., the reward $r_t = R(s_t, a_t)$. For each group we multiply the fraction of admitted students by the expectation of their success. Formally, the reward is defined as follows:

$$R(s_t, \theta_D) = \phi_D \frac{s_t^2 - \theta_D^2}{2s_t} + \phi_A \frac{1 - \theta_A^2}{2}.$$

More details are provided in the appendix.

Dynamics (Transition Function). Given the current state s_t and action a_t, the next state s_{t+1} is determined through the transition dynamics. We assume that exogenous factors as well as the DM's decisions affect the bias state at constant levels $\sigma \in [0,1]$ and $(1-\sigma)$, respectively. In this paper, we mostly focus on two types of deterministic dynamics and their combination:

(1) Representation Dynamics (τ_1): This dynamic reflects the reported positive effect of observing role-models from the same group on academic success [26]. The future bias depends on the current fraction of selected students from the disadvantaged group compared to α. This also corresponds to the notion of demographic parity [3].

$$\tau_1(s_t, \theta_D) = \sigma + (1-\sigma)\frac{s_t - \theta_D}{\alpha s_t}.$$

(2) Relative Success Dynamics (τ_2): This dynamic accounts for the "stigma of incompetence", where group members benefiting from affirmative action are perceived as less capable [25]. The future bias depends on the current success probability of admitted students from the disadvantaged group, compared to that of the advantaged group. This also corresponds to the notion of predictive parity [6] or equality of opportunity [22].

$$\tau_2(s_t, \theta_D) = \sigma + (1-\sigma)\frac{s_t + \theta_D}{1 + \theta_A}.$$

(3) Combined dynamics (τ_3): As both the representation and relative success are likely to affect the bias, this dynamic is the affine combination of the two dynamics τ_1 and τ_2, where the ratio between the two is set by a parameter ρ.

$$\tau_3(s_t, \theta_D) = \rho\tau_1(s_t, \theta_D) + (1-\rho)\tau_2(s_t, \theta_D).$$

Note that all states must be in $[0,1]$, as they represent the upper bound of the success distribution. Thus, the output of a deterministic transition function must be clipped to fit to this range by a minimum and maximum operator, which we omit above for better readability.

Given this MDP, the DM decides on a policy $\pi : S \to A$, which is a mapping from states to actions. We consider only deterministic policies.[2]

[2]For stochastic policies, the policy is a mapping from states to probabilities of selecting each possible action: $\pi(a|s) \in [0,1]$ is the probability that action a is taken for state s when following policy π, $\sum_a \pi(a|s) = 1$. Note that when maximising expected utility of a policy, there always exists a deterministic optimal policy [49].

5 AFFIRMATIVE ACTION AND PREFERENTIAL TREATMENT

In our first experimental analysis, we treat the (deterministic) transition dynamics as fixed parameters of the MDP and show its effect on the actions taken by a **utility-maximising policy**. Here, we are in particular interested in the occurrence of *preferential treatment* and *affirmative actions*.

Affirmative action is defined as an action or a policy favouring individuals belonging to groups regarded as disadvantaged. In this paper, we consider this favouring action to be a lower admission threshold, which is often the use of affirmative action [18]. In this sense, affirmative action is only favoring the disadvantaged group in the current decision / time-step and might not lead to fairer outcomes in the long run.

We consider preferential treatment (PT) as an action or a policy that is likely to lead to greater benefits for the disadvantaged group in the future. This could also include actions that impose harder conditions on the disadvantaged group, e.g., higher admission thresholds, as long as the overall group's welfare will be improved. Here, we assume that a measure that indicates the state of the group's welfare is given, e.g., the societal bias against the group.

We assume all states $s_t \in [\sigma, 1]$ are possible start states and uniformly distributed. For known transition functions, we can use approximate dynamic programming to obtain the utility maximising policy. Using discretisation of the state and action spaces (step size= 0.001), we apply policy iteration for infinite horizon.[3]

In our experiments, the measure of fairness of the resulting population state corresponds to the value of the next bias factor (state), $f_\tau = \tau$. When the bias state equals 1, the success probability of the disadvantaged group members is not impaired, and we see this state as completely fair. The lower the bias factor, the more unfair the state. Hence, this fairness measure corresponds to preferential treatment.

As we can see in Figure 2, for all three types of dynamics there is an interval of high-bias (i.e., low state values) in which no one from the disadvantaged group is admitted because the threshold is equal to the upper bound of their success distribution (i.e., the state). Interestingly, only for the representation dynamic, this is followed by an interval of higher state values in which thresholds for advantaged and disadvantaged group are equal. All dynamics display a clear tipping point (red dashed line) on state values from which on a *preferential treatment* is implemented, i.e., actions that increase future success chances of the whole population by increasing the success chances of the disadvantaged group. Note that, for the relative success dynamic, this actually means increasing the threshold for the disadvantaged group (no affirmative actions), while under representation and the combined dynamics affirmative actions are taken. In particular, relative success and representation dynamics have a contrary effect. For other combinations of the two, by varying ρ, these contrary effects balance out such that the same threshold is set for both groups.

For all dynamics, there exists a small interval of high state values from which the unbiased state can be reached in only one time step (see green separation line in figure 2).

[3]Ties between maximising actions are broken by taking the action that minimises the difference between the two thresholds.

The choice of discount factor γ and model parameters ϕ, σ and α affects the location of the tipping point from where on preferential treatment is enforced. Similarly, they also influence the length of the aforementioned intervals. However, for representation and relative success dynamics, the nature of preferential treatment (affirmative / non-affirmative actions) remains the same and, more importantly, differs between the two dynamics. This underlines the importance of knowing the bias dynamics before employing policies in real-life that are aimed at mitigating societal bias. In this paper we focus on combinations of parameters for which the interval of high state values from which the unbiased state can be reached within one step contains more than a single state.

6 UTILITY-FAIRNESS TRADE-OFF

In this section we analyse the interplay between utility and fairness. Under known dynamics, we examine how a long-term view affects the relation between utility and fairness. Under unknown dynamics, we explore whether and how the level of uncertainty and the choice of policy (realistic/pessimistic) affect the utility-fairness trade-off.

6.1 Utility Maximisation Under Known Dynamics

The following experiment is testing our hypothesis that utility and fairness are aligned under a long horizon view. To do so, we find the utility-maximising policy for each of the dynamics for different discount factors.

We choose the same discount factor γ as used in the utility function as weights for the fairness measure. Furthermore, the considered dynamics are deterministic and any policy considered is deterministic, such that we can omit the expectation in the fairness measure in Equation 2. Since we discretise the state space, we only need to compute a sum instead of the integral. Hence, in our experiments we measure the following policy fairness: $F^\pi = \sum_{s_0 \in S}[\sum_{t=0}^{T} \gamma^t \tau(s_t, a_t)]$.

In Figure 3 we plot the utility and fairness (according to the deployed dynamic) for utility-maximising policies under three dynamics against the value of the discount factor γ they were trained for. The same discount factor γ is used for utility-maximisation and measuring fairness. In plots (a), (c) and (e) we can see utility and fairness measures computed for a time horizon of only 1 step. That is, the utility is the immediate reward and the policy fairness is simply the immediate fairness outcome. For any of the three dynamics, we can tune the utility-fairness trade-off by varying the discount factor: for lower values we get higher rewards with lower fairness, while for larger discount factors we get lower rewards with higher fairness. Yet, when we observe these measures for a longer horizon (utility and fairness measured for 150 steps) in plots (b), (d) and (f), we can see that utility and fairness are aligned and increase with the discount factor.

A short-sighted DM (with lower discount factor) will thus experience the satisfaction of high short-term utility at the cost of low short-term fairness, while both utility and fairness measured over a longer horizon will be low. On the contrary, a far-sighted DM (with higher discount factor) might be unsatisfied with the immediate returns but experiences higher fairness and utility when measured over a longer horizon. Thus, one can conclude that the trade-off

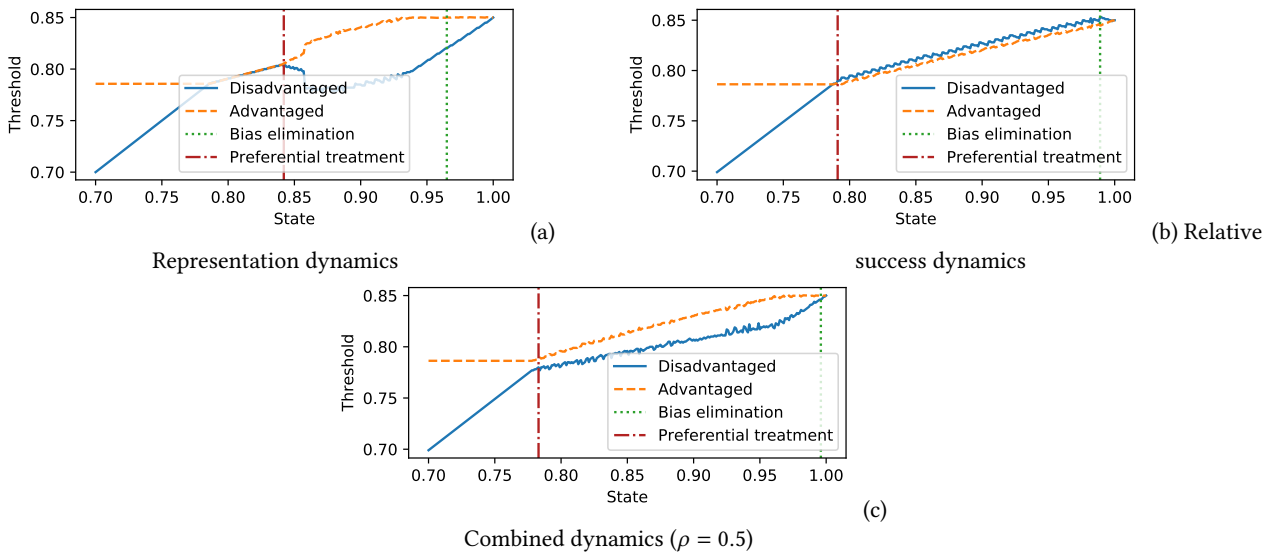


Figure 2: Thresholds θ_A, θ_D of a utility-maximising policy under different dynamics with $\phi = 0.3$, $\sigma = 0.7$, $\alpha = 0.15$, $\gamma = 0.2$.

in this dynamic system is not between fairness and utility, but between measured time horizon and utility. When pushed to increase fairness, both short- and far-sighted DMs will act according to an increased discount factor. However, a more far-sighted DM will do so willingly, while a short-sighted DM will see this as a trade-off for utility.

6.2 Utility Maximisation Under Uncertainty

In the previous sections, we presented specific bias dynamics, or state transition functions, and assumed these to be known by the DM when optimizing their policy. Yet, in reality the true dynamics of the process are seldom known and might even be counter-intuitive. This uncertainty, that is a result of missing information regarding the underlying MDP model, is called epistemic (or parametric) uncertainty [51].[4] We would like to measure both fairness and utility of different uncertainty-aware utility-maximising policies. As before, we assume that all other model parameters are known and constant, including the reward function.

We take the Bayesian approach and assume to have some prior belief ξ_0 over the transition function, and that we can sample the posterior at any point. The transition function can be learned by updating the belief based on observations of the true next state at each step. The uncertainty will then decrease until the transition function is fully learned.

For a given belief distribution ξ over transition functions $\mathbb{T}$, the Bayes-optimal policy $\pi^* \in \Pi$ in some policy class Π can be found by maximising the expected utility of π:

$$\mathbb{E}^{\pi}_{\xi}[U] = \int_{\tau \in \mathbb{T}} U^{\pi}_{\tau} \, \mathrm{d}\xi(\tau).$$

This integral can be approximated by sampling. If we only sample one τ and act according to this, then we obtain the well-known Thompson sampling algorithm. However, for more informed approximation we use backwards induction on multiple MDPs in our experiments as described in Algorithm 1 in [11]. At each time t, we sample n transition functions from the posterior $\tau_1, \ldots, \tau_n \sim \xi_t$. We then calculate an approximately-optimal policy through backwards induction on n MDPs. This is done by maintaining a value function V_i for every τ_i, and at each policy improvement step choosing the action maximising some aggregation of all value functions. The improved policy is defined by $\pi(s) = \mathrm{argmax}_a r(s,a) + \gamma \hat{V}(s,a)$, with aggregated value function $\hat{V}(s,a) = g(V_1(\tau_1((s,a))), \ldots, V_n(\tau_n((s,a))))$. There is no need to aggregate over the rewards because in our case the reward function is known and does not depend on the transition function. The following are two possible aggregation functions.

Mean policy: Aggregate the value functions by taking their mean, such that $\hat{V}(s,a) = \frac{1}{n} \sum_i V_i(\tau_i(s,a))$. This corresponds to the standard Bayesian approach that approximately maximises expected utility.

Pessimistic policy: Aggregate the value functions by taking the minimum over all value functions, such that $\hat{V}(s,a) = min_i\{V_i(\tau_i(s,a))\}$. By taking the minimum over all sampled transition combinations, we obtain an approximate lower bound, and hence a more risk-averse policy.[5]

6.2.1 Experimental Setup. We assume that the transition function τ is a linear function and use Bayesian linear regression [29, 38] to learn it. Given a state s_t, $\tau(s_{t+1}|x_t, A, L) \sim N(Ax_t, L)$, where x_t is a vector including the current state, action and the constant model parameters, A is a coefficient vector and L is the noise prior (variance of the normal distribution N). The mean of the distribution, Ax_t, is the value that maximises the conditional probability of s_{t+1}, so we can get the next state by setting $s_{t+1} = Ax_t$. In our experiments, we define the input of the linear regression to be polynomial

[4] Another kind of uncertainty is aleatoric (or internal) uncertainty, which is a result of inherent stochasticity of the underlying MDP. Yet, in our use case the transition and reward functions are deterministic, so uncertainty of this sort can be ignored.

[5] Note, that while this policy is pessimistic regarding the samples of transition functions, it is still optimistic in assuming that an optimal policy for the worst case sampled transition will be employed in all future time steps.

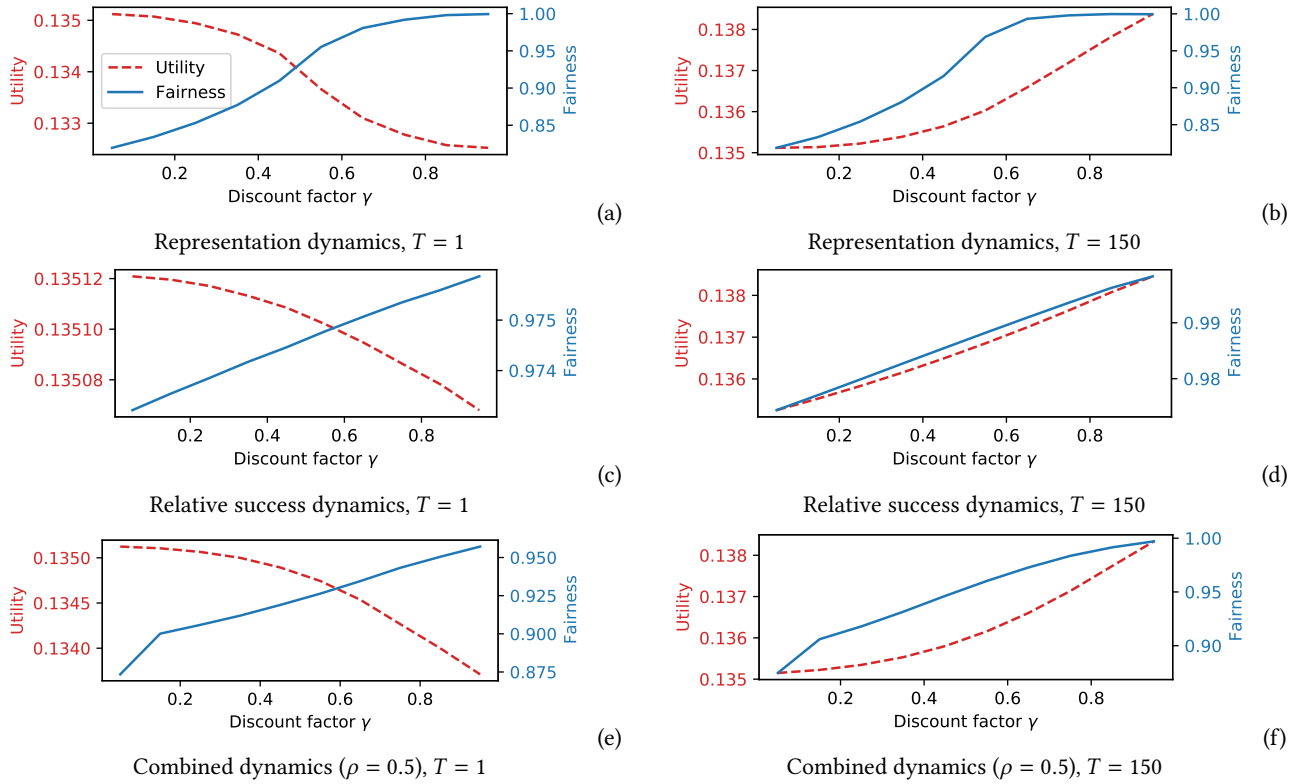


Figure 3: Utility and fairness of utility-maximising policies measured for time horizon T. For $T = 150$, the results are normalised by $1 - \gamma$. Policies are trained for different discount factors γ with $\phi = 0.3, \sigma = 0.7, \alpha = 0.15$.

features of degree 3 of the following values together with their inverse values: the current state s_t, current action $a_t = \theta_D$, $1 + \theta_A$[6], and α the fraction of the population that the DM can admit. Note that other model parameters, such as group sizes, are implicitly included in the calculation of θ_A. Both representation and relative-success dynamics can be represented by this formulation, hence any linear combination of the two can be represented as well. We thus consider this formulation to be expressive enough for our case analysis, where we try to learn the dynamics presented in section 4. As mentioned in Section 4, the output of the (deterministic) transition function, i.e., the next state, needs to be clipped to be between $[0, 1]$. Hence, we only learn from steps within this range.

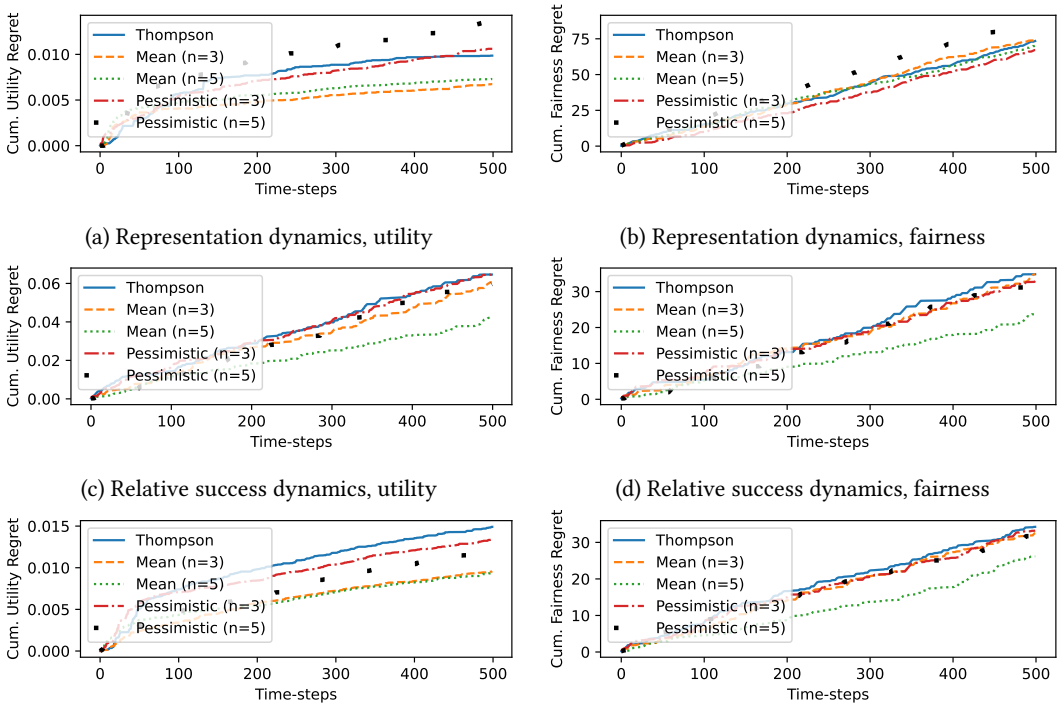
Evaluation Measures. For this setting, we measure both utility and fairness via regret measures. The utility regret is the difference in return between the best possible action and the action that was selected by the policy: $Regret(s, a) = V^*(s) - (r(s, a) + \gamma V^*(\tau(s, a))$, where V^* is the value function for the optimal policy under the true transition function. For fairness regret, we replace the reward with the immediate fairness score and V^* with V^{f*}, which is the value function for the fairness maximising policy under the true transition function. For both regrets, a lower value is preferred.

[6]This parameter is necessary to capture, e.g., the relative success transition function when using a linear regression.

6.3 Experimental Results

Figure 4 shows the cumulative utility regret over time (left) and cumulative fairness regret (right) for the different uncertainty-aware policies when the transition dynamics is unknown. Thompson sampling performs worst in both utility and fairness regrets for relative success and combined dynamics. Only for representation dynamics does Thompson sampling show some advantage, in particular over one of the pessimistic policies. Except for fairness regret for representation dynamics, in all other cases one of the mean policies has the lowest regret. However, among them the performance depends on the sample size, dynamics and regret type. Pessimistic policies perform better than Thompson sampling and worse than mean policies for relative success and combined dynamics. For representation dynamics, the utility regret is worse or equivalent to Thompson sampling, while the fairness regret is the worst for one of the pessimistic policies and is the best for the other.

Notably, the performance order of the policies changes between utility and fairness regret. Based on these results, we can conclude that for a fixed (low) discount factor, uncertainty-aware policies which improve utility regret do not necessarily improve fairness regret when the transition function is unknown. Moreover, the benefit of each policy may change, depending on the underlying dynamics. However, a mean policy seems to be a reasonable choice compared to Thompson sampling and pessimistic policies (depending on the sample size). In the appendix we provide additional



(a) Representation dynamics, utility

(b) Representation dynamics, fairness

(c) Relative success dynamics, utility

(d) Relative success dynamics, fairness

(e) Combined dynamics ($\rho = 0.5$), utility

(f) Combined dynamics ($\rho = 0.5$), fairness

Figure 4: Utility regret and fairness regret of utility-maximising policies, when learning the transition function along 500 **iterations, using discretisation with step size**= 0.01**. Policies are trained for different underlying dynamics with** $\phi = 0.3, \sigma = 0.7, \alpha = 0.15$ **and** $\gamma = 0.2$**.**

experiments, including more samples and a comparison between discount factors.

7 DISCUSSION AND FUTURE WORK

This paper empirically examines utility maximising policies in terms of fairness, under different optimisation choices. Our empirical results indicate that utility and fairness can be aligned in a sequential decision setting. For some dynamics preferential treatment and affirmative actions do not necessarily coincide. For known dynamics, a far-sighted DMs' utility is aligned with fairness. For unknown dynamics, under a fixed (low) discount factor, approaches towards epistemic uncertainty that perform well in terms of utility might perform worse in terms of fairness. Thus, short-sighted DM that are interested in promoting fairness should choose the uncertainty approach based on both utility and fairness, as these choices could lead to better utility-fairness trade-off, even without additional fairness constraints.

Long-Term Fairness. In this paper, we generalise the notion of policy fairness by using a weighted sum of immediate fairness outcomes. DMs face the need to select weights (or discount factor) for utility and policy fairness. Furthermore, we take a consequential approach and do not consider fairness of an action itself. Yet, the end does not justify all means and other fairness notions may be considered when applying fairness interventions. Some actions, such as setting a higher threshold for the disadvantaged group, might be considered as unfair regardless of the long-term effects. Nevertheless, a long-term fairness analysis provides more information to DMs for evaluating possible fairness interventions. We only consider here the effect of one kind of intervention (change of acceptance thresholds), but based on the long-term analysis, the DM could consider other interventions, e.g., increasing the success of admitted students by providing private tutors.

We could also define a fair-state-value function and fair-state-action value function by replacing the rewards with immediate fairness outcomes. This allows to measure the long-term fairness of each state or state-action w.r.t. transition dynamics, policy and discount factor. An analysis based on this fairness measure is left to future work.

Model Extensions. We note that our model has several limitations which could be addressed in future work. First, we assume the DM has full knowledge of the success distributions for both groups. Usually the DM only has access to estimated or predicted success probabilities. These estimations introduce additional errors, biases, feedback effects and uncertainties to the decision making process. Some of these elements were addressed in previous work, but it would be interesting to analyse them combined with epistemic uncertainty. Second, we assume that the transition function is fixed. Our model could be extended to capture cases of changing dynamics as done for non-stationary reinforcement learning settings. In addition, we assume other MDP parameters (reward

function, group sizes, admission vacancies and contribution of exogenous factors) to be known and constant. As these parameters might change over time, our framework could be extended to a belief over the entire MDP, not only the transition function. Lastly, we consider deterministic policies and dynamics, with finite state and action spaces. For our use-case, there was no need for more complex spaces or policy structures, but using parametric policies with gradient-based optimisation might be necessary for other applications. Moreover, the use of stochastic dynamics introduces aleatoric (or internal) uncertainty, which could also be interesting to address when combined with epistemic uncertainty in the future.

ACKNOWLEDGMENTS

This work was supported by the Research Council of Norway under project number 302203.

REFERENCES

[1] Krishna Acharya, Eshwar Ram Arunachaleswaran, Sampath Kannan, Aaron Roth, and Juba Ziani. 2023. Wealth dynamics over generations: Analysis and interventions. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 42–57.

[2] Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, and Kinjal Basu. 2022. Long-term Dynamics of Fairness Intervention in Connection Recommender Systems. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 22–35.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and Machine Learning. fairmlbook. org.

[4] Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Steven Z Wu. 2019. Equal opportunity in online classification with partial feedback. *Advances in Neural Information Processing Systems* 32 (2019).

[5] Marc G Bellemare, Will Dabney, and Rémi Munos. 2017. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*. PMLR, 449–458.

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

[7] P Clavier, S Allassonnière, and E Le Pennec. 2022. Robust Reinforcement Learning with Distributional Risk-averse formulation. *ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA* (2022).

[8] Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. 2020. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*. PMLR, 2185–2195.

[9] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[10] Xolani Dastile, Turgay Celik, and Moshe Potsane. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing* 91 (2020), 106263.

[11] Christos Dimitrakakis. 2011. Robust bayesian reinforcement learning through tight lower bounds. In *European Workshop on Reinforcement Learning*. Springer, 177–188.

[12] Christos Dimitrakakis, Yang Liu, David C Parkes, and Goran Radanovic. 2019. Bayesian fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 509–516.

[13] Hadi Elzayn, Shahin Jabbari, Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, and Zachary Schutzman. 2019. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 170–179.

[14] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*. PMLR, 160–171.

[15] Danielle Ensign, Frielder Sorelle, Neville Scott, Scheidegger Carlos, and Venkatasubramanian Suresh. 2018. Decision making with limited feedback. In *Algorithmic Learning Theory*. PMLR, 359–367.

[16] Hannes Eriksson and Christos Dimitrakakis. 2020. Epistemic risk-sensitive reinforcement learning. *ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 2-4 October 2020* (2020).

[17] European Union. 2021. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Accessed: 2022-08-08.

[18] Hanming Fang and Andrea Moro. 2011. Theories of statistical discrimination and affirmative action: A survey. *Handbook of social economics* 1 (2011), 133–200.

[19] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. 2021. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 445–453.

[20] Ganesh Ghalme, Vineet Nair, Vishakha Patil, and Yilun Zhou. 2021. State-Visitation Fairness in Average-Reward MDPs. *arXiv preprint arXiv:2102.07120* (2021).

[21] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, et al. 2015. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning* 8, 5-6 (2015), 359–483.

[22] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[23] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Advances in Neural Information Processing Systems* 31 (2018).

[24] Hoda Heidari and Jon Kleinberg. 2021. Allocating Opportunities in a Dynamic Model of Intergenerational Mobility. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 15–25.

[25] Madeline E Heilman, Caryn J Block, and Jonathan A Lucas. 1992. Presumed incompetent? Stigmatization and affirmative action efforts. *Journal of Applied Psychology* 77, 4 (1992), 536.

[26] Sarah D Herrmann, Robert Mark Adelman, Jessica E Bodford, Oliver Graudejus, Morris A Okun, and Virginia SY Kwan. 2016. The effects of a female role model on academic performance and persistence of women in STEM courses. *Basic and Applied Social Psychology* 38, 5 (2016), 258–268.

[27] Yaowei Hu and Lu Zhang. 2022. Achieving long-term fairness in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 9549–9557.

[28] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *International conference on machine learning*. PMLR, 1617–1626.

[29] Emilio Jorge, Hannes Eriksson, Christos Dimitrakakis, Debabrota Basu, and Divya Grover. 2020. Inferential induction: A novel framework for bayesian reinforcement learning. (2020).

[30] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in neural information processing systems* 29 (2016).

[31] Nikola H Konstantinov and Christoph Lampert. 2022. Fairness-aware pac learning from corrupted data. *Journal of Machine Learning Research* 23 (2022).

[32] Benjamin Laufer. 2021. Beyond Validity: Current Auditing Methods for Criminal Risk Assessments Do Not Consider Sequential Feedback Effects. (2021).

[33] David Lindner, Hoda Heidari, and Andreas Krause. 2021. Addressing the Long-term Impact of ML Decisions via Policy Regret. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc, 537–544.

[34] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158.

[35] Lydia T Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 381–391.

[36] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[37] Claire Cain Miller. 2015. Can an algorithm hire better than a human. *The New York Times* 25 (2015).

[38] Thomas Minka. 2000. *Bayesian linear regression*. Technical Report. MIT Media Lab. http://research.microsoft.com/en-us/um/people/minka/papers/linear.html

[39] Alan Mishler and Niccolò Dalmasso. 2022. Fair When Trained, Unfair When Deployed: Observable Fairness Measures are Unstable in Performative Prediction Settings. *arXiv preprint arXiv:2202.05049* (2022).

[40] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. 2010. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 799–806.

[41] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. 2010. Parametric return density estimation for reinforcement learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. 368–375.

[42] Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. 2019. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 359–368.

[43] Pegah Nokhiz, Aravinda Kanchana Ruwanpathirana, Neal Patwari, and Suresh Venkatasubramanian. 2021. Precarity: Modeling the Long Term Effects of Compounded Decisions on Individual Instability. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 199–208.

[44] Benjamin Paaßen, Astrid Bunge, Carolin Hainke, Leon Sindelar, and Matthias Vogelsang. 2019. Dynamic fairness-breaking vicious cycles in automatic decision making. *ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019* (2019).

[45] Liam Peet-Pare, Nidhi Hegde, and Alona Fyshe. 2022. Long Term Fairness for Minority Groups via Performative Distributionally Robust Optimization. *ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA* (2022).

[46] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.

[47] José Pombal, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. 2022. Prisoners of Their Own Devices: How Models Induce Data Bias in Performative Prediction. *ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA* (2022).

[48] Bhagyashree Puranik, Upamanyu Madhow, and Ramtin Pedarsani. 2022. A Dynamic Decision-Making Framework Promoting Long-Term Fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 547–556.

[49] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[50] Lydia Reader, Pegah Nokhiz, Cathleen Power, Neal Patwari, Suresh Venkatasubramanian, and Sorelle Friedler. 2022. Models for understanding and quantifying feedback in societal systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1765–1775.

[51] Marc Rigter, Bruno Lacerda, and Nick Hawes. 2021. Risk-averse bayes-adaptive reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 1142–1154.

[52] Samordnaopptak. 2021. The Norwegian Universities and Colleges Admission Service: Gender points. https://www.samordnaopptak.no/info/opptak/opptak-uhg/poengberegning/legge-til-poeng/kjonnspoeng/index.html. Accessed: 2022-08-08.

[53] Sebastian Scher, Simone Kopeinik, Andreas Trügler, and Dominik Kowald. 2022. Long-term dynamics of fairness: understanding the impact of data-driven targeted help on job seekers. *arXiv preprint arXiv:2208.08881* (2022).

[54] Pola Schwöbel and Peter Remmers. 2022. The Long Arc of Fairness: Formalisations and Ethical Discourse. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 2179–2188.

[55] Tareq Si Salem, Georgios Iosifidis, and Giovanni Neglia. 2022. Enabling Long-term Fairness in Dynamic Resource Allocation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 6, 3 (2022), 1–36.

[56] Sean R Sinclair, Siddhartha Banerjee, and Christina Lee Yu. 2022. Sequential fair allocation: Achieving the optimal envy-efficiency tradeoff curve. *ACM SIGMETRICS Performance Evaluation Review* 50, 1 (2022), 95–96.

[57] Sean R Sinclair, Gauri Jain, Siddhartha Banerjee, and Christina Lee Yu. 2020. Sequential fair allocation of limited resources under stochastic demands. *arXiv preprint arXiv:2011.14382* (2020).

[58] Yi Sun, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2022. The Backfire Effects of Fairness Constraints. *ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA* (2022).

[59] Behzad Tabibian, Vicenç Gómez, Abir De, Bernhard Schölkopf, and Manuel Gomez Rodriguez. 2019. Consequential ranking algorithms and long-term welfare. *arXiv preprint arXiv:1905.05305* (2019).

[60] Aviv Tamar, Dotan Di Castro, and Shie Mannor. 2016. Learning the variance of the reward-to-go. *The Journal of Machine Learning Research* 17, 1 (2016), 361–396.

[61] Ruibo Tu, Xueru Zhang, Yang Liu, Hedvig Kjellström, Mingyan Liu, Kun Zhang, and Cheng Zhang. 2020. How Do Fair Decisions Fare in Long-term Qualification?. In *Thirty-fourth Conference on Neural Information Processing Systems*.

[62] Kenneth M Tyler, Falynn A Thompson, Donna E Gay, Jennifer Burris, Howard Lloyd, and Sycarah Fisher. 2016. Internalized stereotypes and academic self-handicapping among Black American male high school students. *Negro Educational Review* 67, 1-4 (2016), 5.

[63] Aline Weber, Blossom Metevier, Yuriy Brun, Philip S Thomas, and Bruno Castro da Silva. 2022. Enforcing Delayed-Impact Fairness Guarantees. *arXiv preprint arXiv:2208.11744* (2022).

[64] Min Wen, Osbert Bastani, and Ufuk Topcu. 2021. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1144–1152.

[65] DJ White. 1988. Mean, variance, and probabilistic criteria in finite Markov decision processes: A review. *Journal of Optimization Theory and Applications* 56, 1 (1988), 1–29.

[66] Joshua Williams and J Zico Kolter. 2019. Dynamic modeling and equilibria in fair decision making. *arXiv preprint arXiv:1911.06837* (2019).

[67] Tongxin Yin, Reilly Raab, Mingyan Liu, and Yang Liu. [n. d.]. Long-Term Fairness with Unknown Dynamics. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

[68] Xueru Zhang and Mingyan Liu. 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control*. Springer, 525–555.