Presented in the Algorithmic Fairness Across Alignment Procedures and Agentic Systems Workshop at ICLR 2026.

LONG-TERM FAIRNESS WITH SELECTIVE LABELS

Giovani Valdrighi
Instituto de Computação
Universidade Estadual de Campinas
Campinas, Brazil

Isabel Valera
Saarland Informatics Campus
Saarland University
Saarbrücken, Germany

Marcos Medeiros Raimundo
Instituto de Computação
Universidade Estadual de Campinas
Campinas, Brazil

ABSTRACT

Long-term fairness algorithms aim to satisfy fairness beyond static and short-term notions by accounting for the dynamics between decision-making policies and population behavior. Most previous approaches evaluate performance and fairness measures from observable features and a label, which is assumed to be fully observed. However, in scenarios such as hiring or lending, the labels (e.g., ability to repay the loan) are *selective labels* as they are only revealed based on positive decisions (e.g., when a loan is granted). In this paper, we study long-term fairness in the selective labels setting, and analytically show that naive solutions do not guarantee fairness. To address this gap, we then introduce a novel framework that leverages both the observed data and a label predictor model to estimate the true fairness measure value by decomposing into the observed fairness and bias from label predictions. This allows us to derive the sufficient conditions to satisfy true fairness from observable quantities by using the confidence in the predictor model. Finally, we rely on our theoretical results to propose a novel reinforcement learning algorithm for effective long-term fair decision-making with selective labels. In semisynthetic environments, the proposed algorithm reached comparable fairness and performance to an agent with oracle access to the true labels.

1 INTRODUCTION

The deployment of machine learning algorithms in critical decision-making scenarios, such as admission processes (Baker & Hawn, 2022; Fuster et al., 2022) and health diagnosis, has motivated the study of algorithmic fairness. One of the most common approaches has been to demand equal benefit from a decision (e.g., acceptance in a process or correct prediction) among different demographic groups of the population (defined by race or gender, for example) (Mehrabi et al., 2021; Angwin et al., 2016). However, Liu et al. (2018) and D'Amour et al. (2020) showed that ensuring fairness at each decision round does not guarantee **fairness in the long-term** due to the feedback loop between policy deployment and the population's reaction. Furthermore, previous decisions determine the available data for policy update, which might mask decisions' unfairness.

In greater detail, long-term fairness considers that individuals are described by features (x_t, z) that relate to a classification label y_t, and both (x_t, y_t) are temporal features dependent on previous actions $a_{i<t}$ and $(x_{i<t}, y_{i<t})$. However, z is a sensitive attribute, such as race or gender (considered binary in this work), and while it can be used to select actions a_t, the expected utility μ of the decision process should be independent of it. The disparity value $|\Delta_t| = |\mu_t^1 - \mu_t^0|$, where μ_t^i is the expected utility for the group i at time t, serves as a measure of the unfairness of the decision process, and algorithms should satisfy that $|\Delta_t| \approx 0$ for every t or when $t \to \infty$. Previous works have considered dealing with this problem with reinforcement learning algorithms (Alamdari et al., 2024; Yu et al., 2022; Yin et al., 2023; Lear & Zhang, 2025; Hu et al., 2023) and optimization approaches when the dynamics model is known (Rateike et al., 2024; Wen et al., 2021). However, these works did not consider a common characteristic: the label y_t is partially observed.

Consider the example of a loan application. The decision-maker has a binary decision to perform (approve or deny), and y_t (payment ability) will only be observed in the case of acceptance. This property, called **selective labels**, presents a great impact on sequential decision-making (Bechavod et al., 2019; Ensign et al., 2018). When considering fairness, the partial observation of labels makes it non-trivial to obtain an unbiased estimate of disparity measures. Our first result (Proposition 3.1) shows that evaluating disparity only on the observed population has no guarantees over the total population.

Motivated by this negative result, in this work, we introduce a general framework for long-term fairness with selective labels, where a decision-maker leverages a label predictor ϕ to perform data imputation. Under this framework, we consider the difference between the true disparity Δ_t of the population and the disparity that the decision-maker sees after data imputation $\tilde{\Delta}_t$. We present a decomposition of the disparity $\tilde{\Delta}_t$ (Theorem 3.1) that relates to Δ_t by the interplay of the rejection rate of groups and the quality of predictions on the rejected population for each group. To present conditions that can be evaluated from the observed data, we introduce generalization bounds of inverse probability weighting to tackle the unknown quality of predictions on the rejected population based on the data of previously accepted individuals. Our main theoretical result (Theo 3.4) presents conditions on the observed disparity $\tilde{\Delta}_t$ and on the bias introduced by the predictor (obtained from generalization bounds) to guarantee low values of true disparity $|\Delta_t|$. Our last contribution is a novel algorithm that learns a policy and predictor model that satisfies fairness in the long-term with access to only observable quantities by satisfying the identified sufficient conditions. Based on the generalization bound, our algorithm incentivizes exploration that improves the stability of the predictor learning. Our proposed algorithm reached long-term fairness comparable to an agent with oracle access to the true disparity measure in semisynthetic environments with high-dimensional features x_t and different fairness notions.

1.1 Related Works

For a more comprehensive discussion on related works, see Appendix B.

Long-Term Fairness Research in long-term algorithmic fairness has primarily leveraged reinforcement learning (RL) and causal modeling. RL solutions included model-based methods(Wen et al., 2021; Rateike et al., 2024; Deng et al., 2024), and adaptations of algorithms such as Q-learning (Alamdari et al., 2024; Chi et al., 2022), RTD3 (Yin et al., 2023) and PPO (Hu et al., 2023; Lear & Zhang, 2025; Yu et al., 2022). Prior work has defined long-term fairness either as minimizing the cumulative disparity over time (Lear & Zhang, 2025; Yu et al., 2022; Yin et al., 2023), or a long-term discounted value function of disparity (Rateike et al., 2024; Hu et al., 2023; Zhang et al., 2020a; Jabbari et al., 2017; Satija et al., 2023; Xu et al., 2024). Puranik et al. (2022) and Raab et al. (2024) introduced population dynamics through time-dependent groups occurrence. Hu & Zhang (2022) leveraged causal modeling to express the temporal dynamics between policy and population. While standard algorithms for constrained decision processes (Achiam et al., 2017; Zhang et al., 2020b) have been evaluated in long-term fairness, Rezaei-Shoshtari et al. (2024) leveraged bisimulation to transform the MDP to a new one where fairness can be satisfied without constraints. Xie & Zhang (2024); Somerstep et al. (2024) considered satisfying fairness over one iteration of strategic classification from individuals. However, all discussed solutions included the assumption that labels are available during learning.

Selective Labels The partial observation of data has been largely studied as selection bias. Its prevalence in standard fairness benchmarks has been highlighted by Fawkes et al. (2024). In decision-making with selective labels, prior work (Kilbertus et al., 2020; Rateike et al., 2022; Keswani et al., 2024; Chang & Wiens, 2024; Frauen et al., 2024; Lakkaraju et al., 2017) has considered an unknown but time-invariant data distribution that is sampled by the agent's policy at each iteration. To avoid exacerbating bias in this setting, (Kilbertus et al., 2020) showed that policies must explore through stochastic actions. Similarly to our approach, (Chang & Wiens, 2024) leveraged a label predictor to assess disparity in the time-invariant distribution setting. More related to this work, (Creager et al., 2020) used a causal estimator to tackle selective labels in dynamic environments. Yet, their analysis was restricted to changes occurring over a single iteration.

2 PRELIMINARIES AND PROBLEM FORMULATION

We consider a decision-making problem where individuals are described by features $x \in \mathcal{X}$ and a binary sensitive attribute $z \in \mathcal{Z} = \{0, 1\}$. Each individual has a latent label $y \in \{0, 1\}$, which is related to their features by the conditional probability $\alpha(x, z) := P(Y = 1 \mid X = x, Z = z)$. For each individual, the decision-maker takes a binary action $a \in \{0, 1\}$ sampled by $\pi(x, z) := P(A = 1 \mid X = x, Z = z)$ where $a = 1$ represents acceptance. In an illustrative scenario of a loan application, X might represent the financial history, Z their race, Y the ability to repay the loan, and A the loan approval decision. For simplicity, we will assume that X is a discrete feature vector; however, our results are also valid for the continuous case. A summary of definitions and notation is presented at Appendix A.

The decision-maker will select actions to maximize a reward function $R(y, a) = a(y - c)$ where $c \in \mathbb{R}^+$ represents the cost of acceptance (e.g., loan amount). Simultaneously, individuals obtain utility from the process by a function, for example, $U(y, a) = 1\{y = a\}$ where $1\{\cdot\}$ is an indicator function. The possible different definitions of R and U reflect the different interests the decision-maker and the applicants might have.

Static Fairness The decision-maker has the objective of maximizing $\mathbb{E}[R(Y, A)]$. However, in high-stakes domains, the employed policy should satisfy that utility is independent of the protected attribute (Barocas et al., 2023). A common approach to evaluate the *static fairness* of a policy is based on the disparity in the expected utility between groups: $\Delta := \mu^1 - \mu^0$, where $\mu^i := \mathbb{E}[U(Y, A)|\mathcal{C}^i]$ is the expected utility of a group i with conditioning event $\mathcal{C}^i$. A fair policy π must satisfy $|\Delta| \leq \omega$, for some small tolerance $\omega \in \mathbb{R}^+$. Different fairness notions can be expressed by the choice of U and $\mathcal{C}$. In this work, we consider three common formulations: 1) *Qualification Parity* (Zhang et al., 2020a) where $\mu^i = \mathbb{E}[Y|Z = i]$; 2) *Accuracy Parity* (Berk et al., 2021) where the utility is the "accuracy" of actions $\mu^i = \mathbb{E}[1\{Y = A\}|Z = i]$; and 3) *Equality of Opportunity* (Hardt et al., 2016) where utility is the true positive rate $\mu^i = \mathbb{E}[A|Y = 1, Z = i]$.

Decision-making induces reactions from the population, where actions have an effect in future states (Perdomo et al., 2020). This motivates the consideration of features (x_t, y_t) as time-dependent, commonly employing the Markov Decision Process formulation (Gohar et al., 2024).

Definition 1 (Markov Decision Process (MDP) adapted from Wen et al. (2021))**.** A Markov Decision Process (MDP) is a tuple $\mathcal{M} := \langle S, \mathcal{A}, P_0, P_\mathcal{T}, R\rangle$ where S is a set of states, $\mathcal{A}$ is a set of actions, $P_0 : S \to [0, 1]$ is the initial distribution of states, $P_\mathcal{T}(s, a, s') : S \times \mathcal{A} \times S \to [0, 1]$ is the probability of reaching state s' given action a at state s, $R : S \times \mathcal{A} \to \mathbb{R}$ is a reward function.

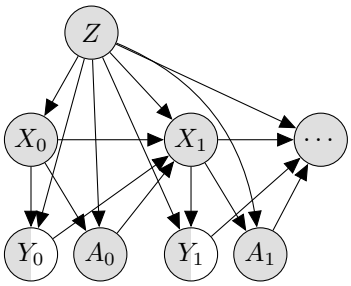


Figure 1: Graphical model for $\mathcal{F}$-MDP. Y_t is partially observed depending on $A_t = 1$.

We extended the MDP definition to represent the dynamic process of decision-making by defining the state as the observable features of each individual and the transition and reward functions as dependent on the binary label.

Definition 2 ($\mathcal{F}$-MDP)**.** An $\mathcal{F}$-MDP is a tuple $\langle S, \mathcal{A}, P_0, P_\mathcal{T}, R, U, \alpha\rangle$ where $\langle S, \mathcal{A}, P_0, P_\mathcal{T}, R\rangle$ follows the MDP definition with $S = \mathcal{X} \times \mathcal{Z}$ and $\mathcal{A} = \{0, 1\}$. Furthermore, each state has an associated label y_t that follows the distribution $Y_t|X_t, Z \sim Be(\alpha(X_t, Z))$. R and U are the reward and utility functions, respectively, evaluated with tuples (y_t, a_t).

This definition reorganizes the variables of the introductory decision-making problem into a dynamic environment (Fig. 1). In this model, the individuals' sensitive attribute Z and α are assumed to be time-invariant, similarly to previous works (Rateike et al., 2024; Hu & Zhang, 2022). Furthermore, we consider that the action A is independent of Y when X, Z are known. While there is no direct effect of action a_t on the label y_t, it will influence future labels by the path $A_t \to X_{t+1} \to Y_{t+1}$.

Long-term Fairness In a dynamic $\mathcal{F}$-MDP, fairness constraints must be satisfied at each step. We are interested in the per-step disparity $\Delta_t = \mu_t^1 - \mu_t^0$, where $\mu_t^i := \mathbb{E}[U(Y_t, A_t) \mid \mathcal{C}^i]$. The ex-

pectation is taken over the distribution induced by policy π and environment dynamics $(P_0, P_{\mathcal{T}}, \alpha)$, conditioned on $\mathcal{C}^i := \{Z = i\}$ for qualification/accuracy parity or $\mathcal{C}^i := \{A = 1, Z = i\}$ for equality of opportunity. The optimization problem is then:

$$\max_{\pi} \quad \mathop{\mathbb{E}}_{\pi,\alpha,P_{\mathcal{T}},P_0}\left[\sum_{t=1}^{T} R(Y_t, A_t)\right] \qquad \text{s.t.} \qquad |\Delta_t| \leq \omega \ \forall t \tag{1}$$

This framing considers that the decision-making can steer the population towards future states where achieving fairness might have a lower cost to the reward objective.

Selective Labels In many practical settings, while the decision-maker observes (X_t, Z) and perform decisions based on it, the true label Y_t is hidden and observed only for accepted individuals $(A_t = 1)$[1]. However, as previously discussed, the information of Y_t is necessary for evaluating the group-wise utilities. The reward value of an action $A_t = 0$ is always 0, independently of Y, however, for example, with $U(y, 0) = 1\{y = 0\}$, the utility could be either 1 or 0 if $A_t = 0$ depending on the unknown label y. This creates a challenge: how to evaluate fairness metrics that depend on the label Y_t?

3 Measuring Disparity Under Selective Labels

In this section, we study how we can employ a label predictor and quantities calculated from observed data to constrain true disparity under the selective labels scenario. Initially, we discuss the pitfalls of a simpler solution to calculate disparity.

Disparity in the Accepted Population Under the selective labels scenario, the decision-maker must evaluate fairness by only using labels Y_t from previously accepted individuals. A naive approach is to compute the disparity only with the accepted subset of the population as follows:

$$\Delta_t^{A=1} = \mathbb{E}[U(Y_t, A_t) \mid \mathcal{C}^1, \{A_t = 1\}] - \mathbb{E}[U(Y_t, A_t) \mid \mathcal{C}^0, \{A_t = 1\}] \tag{2}$$

However, this measure is unaware of the disparity present within the rejected population.

Proposition 3.1 (Formal presentation in Appendix C.1)**.** *For the three fairness notions (Sec. 2), $\Delta_t^{A=1} = 0$ is not a sufficient condition to have $\Delta_t = 0$. In particular, for equality of opportunity, $\Delta_t^{A=1}$ is always* 0.

Proposition 3.1 shows that $\Delta_t^{A=1}$ is a flawed objective for learning when disparity measures are dependent on Y_t. A policy can be optimized to minimize $\Delta_t^{A=1}$ and learn to mask the true disparity measure. For instance, with qualification parity, the policy can learn to accept individuals who have a similar distribution of labels between groups without ensuring equal qualification in the total population.

Imputation Model To be able to calculate disparity from the complete population, the decision-maker can employ a model $\phi : \mathcal{X} \times \mathcal{Z} \to [0, 1]$ to predict unseen labels and evaluate fairness based on the imputed labels. We set predicted labels sampled by $\hat{Y}_t | X_t, Z \sim Bernoulli(\phi(X_t, Z))$, and define the imputed label as $\tilde{Y}_t = A_t Y_t + (1 - A_t)\hat{Y}_t$. That is, with acceptance $(A_t = 1)$ the true label is used (Y_t) and with rejection $(A_t = 0)$, we use the prediction $(\hat{Y}_t)$. We then compute the observed disparity $\tilde{\Delta}_t = \tilde{\mu}_t^1 - \tilde{\mu}_t^0$, where $\tilde{\mu}_t^i := \mathbb{E}[U(\tilde{Y}_t, A_t)|\tilde{\mathcal{C}}^i]$[2] is the utility calculated using $\tilde{Y}_t$.

Due to the complexity of real-world data, predictions will not correctly classify all samples and can amplify biases due to the data availability. Following, we analyze the relation between errors from the predictor model and the distortion of true fairness.

[1]Our framework is capable of handling two formulations of selective labels: 1) the label Y is hidden but exists for every individual and 2) the label Y is only realized with the acceptance and is undefined for rejected individuals.

[2]With equality of opportunity, the condition $\mathcal{C}^i = \{Z = i, Y = 1\}$ is replaced by $\tilde{\mathcal{C}}^i = \{Z = i, \tilde{Y} = 1\}$.

3.1 Decomposition of Disparity with a Label Predictor

As discussed by previous works, the policy influences disparity by two paths: the direct influence from decisions at each iteration and the indirect influence from previous decisions that determined the current state (Lear & Zhang, 2025; Hu & Zhang, 2022). With our imputation model, the policy π has an extra effect on the observed disparity $\tilde{\Delta}_t$: it sets when the predictor ϕ is used for data imputation. We formalize this effect in the following theorem.

Theorem 3.1 (Observed Disparity Decomposition)**.** *Let $\epsilon_t^i := \mathbb{E}[\hat{Y}_t - Y_t | A_t = 0, Z = i]$ and $r_t^i := P(A_t = 0 | Z = i)$ be, respectively, the predictor error on the rejected population and the rejection rate for group i at time t. Then, the observed disparity $\tilde{\Delta}_t$ can be decomposed for each fairness notion:*

- *Qualification parity* ($\tilde{\Delta}_t = \mathbb{E}[\tilde{Y}_t | Z = 1] - \mathbb{E}[\tilde{Y}_t | Z = 0]$)*:* $\tilde{\Delta}_t = \Delta_t + (r_t^1 \epsilon_t^1 - r_t^0 \epsilon_t^0)$
- *Accuracy parity* ($\tilde{\Delta}_t = \mathbb{E}[1\{\tilde{Y}_t = A_t\} | Z = 1] - \mathbb{E}[1\{\tilde{Y}_t = A_t\} | Z = 0]$)*:* $\tilde{\Delta}_t = \Delta_t - (r_t^1 \epsilon_t^1 - r_t^0 \epsilon_t^0)$
- *Equality of opportunity* ($\tilde{\Delta}_t = \mathbb{E}[A_t | Z = 1, \tilde{Y}_t = 1] - \mathbb{E}[A_t | Z = 0, \tilde{Y}_t = 1]$)*:* $\tilde{\Delta}_t = \mu_t^1 \kappa_t^1 - \mu_t^0 \kappa_t^0$ *where* $\mu_t^i = \mathbb{E}[A_t | Z = i, Y_t = 1]$*,* $\kappa_t^i = 1 - r_t^i \epsilon_t^i / \tilde{\phi}_t^i$*, and* $\tilde{\phi}_t^i = P(\tilde{Y}_t = 1 | Z = i)$*.*

Theorem 3.1 shows that the observed disparity $\tilde{\Delta}_t$ is confounded by the bias on $r_t^i \epsilon_t^i$ (or $r_t^i \epsilon_t^i / \tilde{\phi}_t^i$) that relates the policy rejection rate r_t^i with the predictor error ϵ_t^i. When optimizing for fairness using observed data, an algorithm might inadvertently exploit the imputation bias by reducing $|\tilde{\Delta}_t|$ without improvements in $|\Delta_t|$. To avoid this, the decision-maker could obtain a bounded value of true disparity $|\Delta_t|$ by balancing the rejection rate and group error, as we show in our next result.

Theorem 3.2 (Sufficient Conditions for Bounding True Disparity)**.** *For each fairness notion and a constant $\omega \in \mathbb{R}^+$, the following conditions are sufficient to bound the true disparity $|\Delta_t| \leq \omega$:*

- *Qualification parity and accuracy parity:* $|(r_t^1 \epsilon_t^1 - r_t^0 \epsilon_t^0)| \leq \omega/2$ *and* $|\tilde{\Delta}_t| \leq \omega/2$*.*
- *Equality of opportunity:* $|(r_t^1 \epsilon_t^1 / \tilde{\phi}_t^1 - r_t^0 \epsilon_t^0 / \tilde{\phi}_t^0)| \leq (1 - v_t)\omega/2$ *and* $|\tilde{\Delta}_t| \leq (1 - v_t)\omega/2$ *where* $v_t := \max_i r_t^i \epsilon_t^i / \tilde{\phi}_t^i$*.*

This theorem shows that to be able to constrain the true disparity with an upper bound of ω, the uncertainty induced by the predictor error demands that the observed disparity satisfy an even lower upper bound $\omega/2$. Similarly, the imputation bias should also be constrained by $\omega/2$. The conditions for equality of opportunity are stricter when v_t gets closer to 1.

However, conditions from Theorem 3.2 are not actionable, as they depend on the error in the rejected population, which has unobserved labels. To make these conditions practical, in the following section, we leverage the theory of domain adaptation to bound the error on the rejected population.

3.2 Bounding True Disparity from Observable Quantities

In an iterative learning process, the decision-maker will employ a sequence of policies $\pi[1], \ldots, \pi[K]$ for K iterations to perform actions and select clients. Thus, this (labeled) collected data from previous iterations can be used to estimate the error of the (unlabeled) rejected population using the theory of domain adaptation. For simplicity, we will omit the subscript t in this section.

Let $A[k] \sim \pi[k]$ be the decision at iteration k. The feature distribution for individuals in group i rejected by the current policy $\pi[K]$ is $D_R^i(x) := P(X = x | A[K] = 0, Z = i)$, and for those accepted in any iteration up to K is $D_A^i(x) := P(X = x | \bigvee_{k=1}^K A[k] = 1, Z = i)$. By defining the error function $\epsilon(x, i) = \mathbb{E}[\hat{Y} - Y | X = x, Z = i]$, the error over the rejected population is $\epsilon^i = \mathbb{E}_{X \sim D_R^i}[\epsilon(X, i)]$. We can estimate this error using the accepted data via Inverse Propensity Weighting (IPW), in which from a random set of N^i samples collected with D_A^i we can estimate ϵ^i by $\hat{\epsilon}_{A,\mathrm{w}}^i = \sum_{j=1}^{N^i} \epsilon(x_j, i) \mathrm{w}(x_j, i)$ where $a[1:K]^i = P\left(\bigvee_{k=1}^K A[k] = 1 | Z = i\right)$ is the acceptance

rate up to iteration K and:

$$\mathrm{w}(x,i) = \frac{D_R^i(x)}{D_A^i(x)} = \frac{a[1:K]^i}{r^i} \cdot \frac{1-\pi[K](x,i)}{1-\prod_{k=1}^{K}(1-\pi[k](x,i))} \tag{3}$$

The weight $\mathrm{w}(x,i)$ quantifies how much more likely the features x for group i are to be found in the rejected population relative to the accepted population. IPW is known for suffering from high variance whenever the denominator approaches 0 (Rateike et al., 2022); however, our approach mitigates this issue by defining D_A^i as the combination of all K policies obtained during training. Furthermore, we leverage generalization bounds on the IPW estimator based on the variance from Cortes et al. (2010) to present an upper bound of the error.

Assumption 1 (Overlap)**.** For each group i, D_R^i is absolutely continuous with respect to D_A^i.

Theorem 3.3 (Adapted from (Cortes et al., 2010))**.** *Let $d < \infty$ be the pseudo-dimension of the hypothesis space of predictor models ϕ and N^i be the number of accepted samples for group i, the error on the rejected population ϵ^i for group i is bounded by $\overline{\epsilon}^i$ with high probability:*

$$\epsilon^i \leq \hat{\epsilon}^i_{A,\mathrm{w}} + \mathcal{O}\left(\sqrt{d_2(D_R^i||D_A^i)}/\sqrt{N^i}\right) := \overline{\epsilon}^i \tag{4}$$

where $d_2(D_R^i||D_A^i) = \mathbb{E}_{D_A^i}\left[\mathrm{w}(x,i)^2\right]$ is the Renyi divergence with factor 2.

This bound permits us to be explicit about the quality of the IPW estimator of the error, which depends on the distance between distributions of rejected and accepted individuals and the number of samples N^i. The bound will get tighter when more data is collected and when the policy is less strict in the separation between rejected and accepted individuals. By substituting this error bound $\overline{\epsilon}^i$ in our framework, we arrive at our main practical result: *a set of fully observable and enforceable conditions for guaranteeing long-term fairness.*

Theorem 3.4. *For each fairness notion and a given constant $\omega \in \mathbb{R}^+$, the following conditions are sufficient to have $|\Delta| \leq \omega$ with high probability:*

- *Qualification parity and accuracy parity: $\sum_i r^i|\overline{\epsilon}^i| \leq \omega/2$, and $|\tilde{\Delta}| \leq \omega/2$.*
- *Equality of opportunity: $\sum_i r^i|\overline{\epsilon}^i|/\tilde{\phi}^i \leq (1-v)\omega/2$ and $|\tilde{\Delta}| \leq (1-v)\omega/2$, where $v = \max(r^i|\overline{\epsilon}^i|/\tilde{\phi}^i)$.*

This final theorem presents practical conditions to satisfy true fairness. It shows that an algorithm that reaches observed fairness in $\tilde{\Delta}$ can ensure true fairness Δ by two paths: 1) reduce the error bound $\overline{\epsilon}^i$ (by reducing $\hat{\epsilon}^i_{A,\mathrm{w}}$ or reducing the separation between accepted and rejected distributions) or 2) reduce the rejection rate r^i of the policy π for groups with high error bound, therefore reducing the reliance on imperfect predictions for that group.

4 METHOD

We present an algorithm for **SE**lective **L**abels in **L**ong-term **F**airness (SELLF) that optimizes the policy π with regularization based on estimates of a predictor ϕ and promotes actions that ensure higher confidence in its estimates. We introduce a new loss term in the PPO algorithm (Schulman et al., 2017) and utilize the advantage regularization approach in (Yu et al., 2022) to constrain the policy. Simultaneously, the predictor model is learned with the data collected by PPO using IPW.

PPO is a policy gradient method for reinforcement learning capable of handling continuous state spaces. Defining the value of a state $V(s) = \mathbb{E}[\sum_t^T R(Y_t, A_t)|S_0 = s]$ and the q-value of a state, action pair $Q(s,a) = \mathbb{E}[\sum_{t=1}^T R(Y_t, A_t)|S_0 = s, A_0 = a]$, with both quantities reflecting the long-term returns, the advantage function is $A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$. One of the main contributions of PPO is the clipping of the advantage to prevent gradient steps from moving the policy further away from the one from which data was collected. It uses the objective:

$$L^{PPO} = \mathbb{E}[\min(r_t(\theta_\pi)A(s_t,a_t), \mathrm{clip}(r_t(\theta_\pi), 1-\epsilon, 1+\epsilon)A(s_t,a_t))] \tag{5}$$

where $r_t(\theta_\pi) = \pi(s_t)/\pi_{\text{old}}(s_t)$ sets the importance of each sample and ϵ is a clipping parameter. Furthermore, a neural network is used to approximate the value function V. We use the approach of advantage regularization introduced by Yu et al. (2022) to satisfy $|\tilde{\Delta}| \leq \omega/2$. The advantage function is penalized as $\hat{A}_\beta(s_t, a_t) = \hat{A}(s_t, a_t) - \beta_1 \max\{|\tilde{\Delta}_t| - \omega/2, 0\}$ with β_1 as a penalization weight. In particular, with qualification parity, we alter the penalization procedure to be based on $|\tilde{\Delta}_{t+1}|$ (replacing $\tilde{\Delta}_t$ with $\tilde{\Delta}_{t+1}$) as an action does not influence the disparity of the current iteration. The advantage will be reduced whenever $|\tilde{\Delta}_t| \geq \omega/2$.

According to Theorem 3.4, constraining $|\Delta_t|$ requires us to reduce the term $r_t^i|\bar{\epsilon}_t^i|$ (or $r_t^i|\bar{\epsilon}_t^i|/\tilde{\phi}_t^i$). To minimize it, we leverage the upper bound of $\bar{\epsilon}_t^i$ (Theorem 3.3)[3]. While the predictor error will be optimized with ϕ, the Renyi divergence term is a function of π. If the divergence becomes large, the bound on $\bar{\epsilon}_t^i$ becomes loose, the propensity weights increase, and our primary goal of reducing $|\Delta_t|$ is no longer guaranteed. This motivates using the Renyi divergence as a regularization term L^{Renyi} in the combined learning objective $J(\theta_\pi) = L^{PPO} + \beta_2 L^{Renyi}$ where $c_t^i = r_t^i$ ($c_t^i = r_t^i/\tilde{\phi}_t^i$ for equality of opportunity) and:

$$L^{Renyi} = c_t^1 \hat{\mathbb{E}}[\mathrm{w}(x_t, 1)^2 | Z = 1] + c_t^0 \hat{\mathbb{E}}[\mathrm{w}(x_t, 0)^2 | Z = 0] \tag{6}$$

Moreover, we leverage data collected by PPO to train the predictor ϕ with binary cross-entropy loss. We employ inverse propensity weighting to adjust the distribution of samples that were collected under a selection bias imposed by π. That is, the predictor ϕ is optimized to minimize:

$$L^{Classif} = \textstyle\sum_{i \in \{0,1\}} \mathbb{E}_{D_A^i}[\mathrm{w}(x_t, i)\ell(y_t, \phi(x_t, i))/\mathrm{w}(i)] \tag{7}$$

where ℓ is the binary cross-entropy evaluated at each sample and the weights $\mathrm{w}(x_t, i)$ (Eq. 3) shift the distribution to the overall distribution of individuals. To tackle the variance of IPW, we include the normalization term $\mathrm{w}(i) = \sum_{z=i} \mathrm{w}(x_t, z)$ that is used in self-normalized IPW (Swaminathan & Joachims, 2015). The pseudocode for SELLF is presented in Appendix D.

5 Experiments

We evaluated SELLF in semisynthetic environments, performing an ablation study of our solution and a comparison to baselines. We define the initial distribution of individuals based on real-world data and assume simple dynamics in three different environments: a loan application based on FICO scores (Liu et al., 2018), a crime recidivism based on COMPAS (ProPublica) (at Appendix H.4), and a new environment that simulates school admission based on ENEM (INEP, 2025) (Brazilian high school exam). Our experiments are limited to semisynthetic datasets as historical data cannot be used to assess the effect of decision interventions, and real-world evaluations are unfeasible. Our experiments considered a variety of base distributions and simulated dynamics to strengthen our results.

Simulation To simulate the $\mathcal{F}$-MDP, we define the distributions P_Z, P_0, $P_{Y_t|X_t,Z}$, $P_\mathcal{T}$, and create a pool of individuals that follow the joint distribution. At each iteration, given a sampled individual (z, x_t, y_t) from the pool, the decision a_t is sampled from $\pi(x_t, z)$. With (y_t, a_t), we calculate the reward and update the feature x_{t+1} according to the modeled transition $P_\mathcal{T}$ and return this individual to the pool. This procedure induces the update of $P_{X_t|Z}$ to $P_{X_{t+1}|Z}$. For a detailed description of how probabilities were defined based on real datasets for each setting, we refer to Appendix E. Each agent starts with a resource of 1,000, which is updated based on obtained rewards.

Baselines We compare the proposed algorithm SELLF with PPO, designed to maximize reward, two long-term fairness RL algorithms, POCAR (Yu et al., 2022) and ELBERT (Xu et al., 2024), and a constrained RL algorithm, FOCOPS Zhang et al. (2020b). For the baseline algorithms that do not consider the partial observation of features Y, we calculate the necessary metrics using the accepted population, similar to $\Delta^{A=1}$ (Eq. 2). We also implemented a variation of POCAR, which has oracle access to the true disparity Δ, and thus serves as a reference of the attainable fairness without selective labels. To showcase the capabilities of SELLF in more realistic settings, where

[3] An analysis of each term in the bound from Theorem 3.3 during learning is presented in Appendix H.1.

AS. 1 does not hold, we define SELLF (Semi-stoc.) that rejects individuals if $\pi(x, i) < 0.25$ and samples $A_t \sim Bernoulli(\pi(X_t, i))$ if $\pi(x, i) \geq 0.25$.

Experimental Settings Algorithms were trained for 500,000 environment steps. Hyperparameters from PPO, which are common to all tested methods, were adopted from Yu et al. (2022). The disparity constraint was set to $\omega = 0.05$, and fairness-specific hyperparameters were tuned for each algorithm. We report results from the hyperparameter configuration that achieved the highest reward while satisfying disparity constraints. If no configuration satisfied the constraints, we report the one with the lowest disparity. Appendix F presents a complete description of the experimental procedure. In this section, we present results with a linear predictor ϕ, and similar results with complex predictors ϕ are presented in Appendix H.2.

5.1 Lending Environment

We consider a simulated lending environment where each individual is described by a credit score $x_t \in \{1, 2, \ldots, 10\}$, with higher scores having a higher probability of repayment. At each timestep, the decision-maker can either approve or reject a loan application. If rejected, the individual's score remains unchanged. If approved, the score increases by one upon repayment ($y_t = 1$) or decreases by one upon default ($y_t = 0$). We set the cost of acceptance as $c = 0.8$, motivated by the high cost of false positives (defaults) in lending applications.

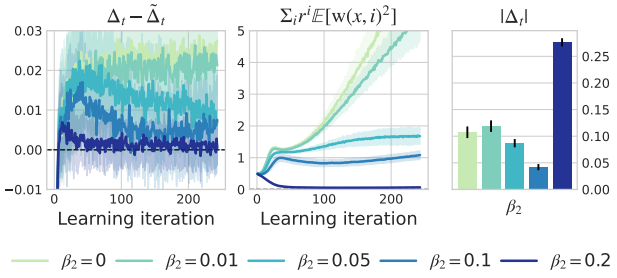


Figure 2: SELLF algorithm executed in the lending environment with $\beta_1 = 5$ and varying values of β_2. We display measures during learning and the disparity of the final policy. Results are averaged with 25 repetitions.

Ablation Study We analyze the effect of the Renyi regularization (Eq. 6) on SELLF by varying the weight β_2. Using the accuracy parity fairness notion, we fixed $\beta_1 = 5$ (weight of $|\tilde{\Delta}_t|$ penalization) and evaluated $\beta_2 \in \{0, 0.01, 0.05, 0.1, 0.2\}$. Fig. 2 displays the behavior during learning of the gap between true and observed disparity, the Renyi regularization, and the final true disparity achieved by the trained agent. For values of $\beta_2 < 0.1$, the disparity gap increased during the initial training phase, ending with values higher than 0.01. Similarly, the Renyi regularization drastically increases over time for these values of β_2. In contrast, with $\beta_2 = 0.1$ and $\beta_2 = 0.2$, the disparity gap is minimized, going to 0 as training progresses. $\beta_2 = 0.1$ also presented the lowest true disparity value among all configurations. A large weight, such as $\beta_2 = 0.2$, can guide the policy to over-accept, which can present high disparity with the accuracy parity fairness notion whenever groups are not equally qualified. For that reason, $\beta_2 = 0.2$ presented the highest true disparity. This study confirmed the importance of the Renyi regularization and demonstrated that with a tuned hyperparameter, we can reach improvements in long-term fairness with selective labels. In our experiments (Appendix F), $\beta_1 = 1$, $\beta_2 = 0.05$ were identified as optimal in most of the environments.

Comparative Results We compare SELLF with baseline algorithms with equality of opportunity. Fig. 3 displays the behavior of trained agents for 10,000 iterations in the environment. PPO, POCAR, ELBERT, and FOCOPS reached comparable rewards, despite only PPO being designed to maximize reward. However, all four agents resulted in disparity above 0.3. As we showed in Proposition 3.1, $\Delta^{A=1}$ is a flawed objective and always 0 for equality of opportunity. For that reason, fairness-aware algorithms behave like PPO, being unable to address unfairness in the total population. SELLF and POCAR (Oracle) obtained the same disparity of 0.05 during the observed period. However, SELLF was able to obtain a higher cumulative reward. This occurs as SELLF presented a higher acceptance rate than POCAR (Oracle), willing to accept individuals with higher risk to reduce

the separation between accepted and rejected populations. In this setting, while SELLF (Semi-sto.) presented a higher disparity than SELLF, it was still below the constraint of 0.05.

5.2 School Admission Environment

Our school admission environment is inspired by the ENEM, a Brazilian national exam. At each timestep t, the decision-maker selects individuals for a preparatory program and can assess the performance on the exam y_t (pass/not pass) of accepted ones, while the remaining labels are unobserved. The environment dynamics are as follows: a student's probability of passing the next exam ($y_{t+1} = 1$) increases if they are selected ($a_t = 1$) or pass the current exam ($y_t = 1$). Furthermore, there is a decrease in the probability of passing the exam between timesteps due to the effect of age, which is present independently of the decision. The conditioned distribution $Y_t \mid X_t, Z$ is a logistic regression learned from data, with X_t having 126 dimensions. The cost is set as $c = 0.5$. In this study, we perform experiments using the qualification parity fairness notion.

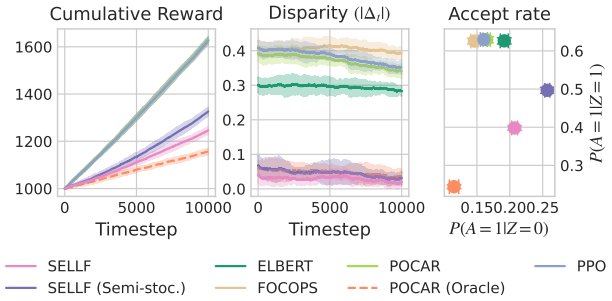


Figure 3: Reward and true disparity (equality of opportunity) over time obtained by optimized agents in the lending environment. Results are obtained with 10 repetitions; confidence intervals are displayed, but are tight. SELLF can ensure the same fairness as the baseline with oracle access and a higher reward.

Comparative Results Fig. 4 displays the result of the trained agents. PPO, ELBERT, POCAR, and FOCOPS obtained the highest rewards. However, four algorithms ended with a disparity higher than the initial value of 0.125. This shows that by optimizing $\Delta^{A=1}$, an agent can even cause harm in the long-term. SELLF, SELLF (Semi-sto.) and POCAR (Oracle) presented a reduction in disparity; however, not being able to reach values lower than $\omega = 0.05$. As the qualification of individuals is highly influenced by the initial state and transition dynamics, agents have less effect on it. All algorithms resulted in an increase in the qualification of the unprivileged group over time, with the highest increase obtained by SELLF.

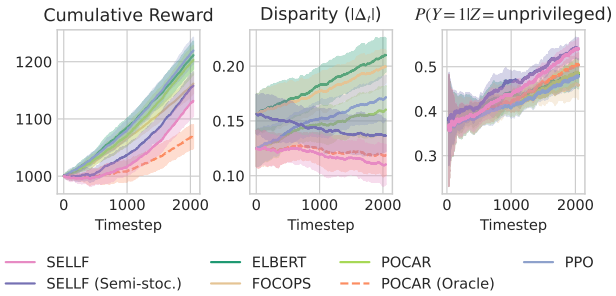


Figure 4: Reward and true disparity (qualification) obtained in the school admission environment. Results are obtained with 10 repetitions.

6 Discussion

Assumptions Our theoretical analysis relies on two simplifying assumptions. First, the $\mathcal{F}$-MDP assumes stationary group dynamics. While this may not hold over extended periods, in practice, the model could be periodically retrained to adapt to new dynamics. Second, our error bounds assume overlap between the distributions of rejected and accepted individuals. That is, every individual who has a non-zero probability of being rejected also has a non-zero probability of being previously accepted. This requirement is consistent with the need for active exploration; the decision-maker must sometimes accept uncertain applicants to gather data and prevent convergence to a suboptimal policy, a principle argued by Kilbertus et al. (2020).

Dependence on IPW As previously discussed, IPW can introduce high variance and learning instability if action probabilities become too small (Swaminathan & Joachims, 2015). While SELLF uses the importance weights in the Renyi and classification losses, our solution presents two safeguards to obtain reduced variance. First, the importance weights are calculated by the aggregated probability of actions from all previous policies. This cumulative probability provides a more stable denominator, preventing it from approaching zero. Second, the Renyi loss objective itself incentivizes the policy to reduce the magnitudes of weights, as demonstrated in Appendix H.3.

Low-exploration Regimes In low-exploration regimes, where an optimal policy has a low acceptance rate, selective labels severely limit the data available for estimating group utilities. In this setting, the Renyi regularization present in SELLF will increase the acceptance rate by approximating the accepted and rejected population. This will distance the optimized policy from the optimal one, with the objective of improving estimates of group utilities. However, the algorithm remains effective; for example, in the lending environment (Sec. 5.1) with acceptance cost of $c = 0.8$, SELLF was able to obtain a positive reward over the deployment period.

7 Conclusion

We studied the problem of long-term fairness under selective labels. In this scenario, the decision-maker must maximize reward while satisfying fairness in regard to labels, which are only observed in the case of acceptance. We present a modeling framework based on an MDP where a predictor model is used to infer unseen labels. Under this new configuration, we presented a theoretical analysis of the relation between true and observed disparity, which was then used to motivate our proposed algorithm. By leveraging the estimates of unfairness obtained by the predictor model and a confidence bound on these estimates, we introduce a simple and flexible reinforcement-learning algorithm. In two semisynthetic environments, our algorithm presented the highest improvements in fairness, reaching similar results to an agent with oracle access to labels. Future work includes the adaptation of our theoretical results to an offline algorithm that leverages historical data, as in highly consequential settings, deploying a policy for learning might be infeasible. Furthermore, future directions also include the study of the setting in which the decision-maker selects an action among multiple possibilities (non-binary), each with different effects.

Acknowledgments

This project was supported by the brazilian Ministry of Science, Technology and Innovations, with resources from Law nº 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Arquitetura Cognitiva (Phase 3), DOU 01245.003479/2024 -10. This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brazil. Process Number #2024/17292-9.

ETHICAL STATEMENT

The presented research encompasses topics of fair application of machine learning in social contexts. To evaluate our proposed algorithm, real-world data was used with sensitive information such as race and gender. Both datasets were anonymized by the source with no identifiable information available.

LLM Usage The authors acknowledge the use of LLM-based tools (Gemini) as a writing assistant to improve text clarity.

REPRODUCIBILITY STATEMENT

All theoretical results presented in the main paper have their proof presented in the Appendix C. The code with the implementation of all algorithms and experiments is included as a supplemental material, with instructions on how to reproduce results. Datasets employed are open and can be downloaded from the references, and the preprocessing steps are presented in Appendix E and in the included code.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pp. 22–31. PMLR, 2017. URL https://dl.acm.org/doi/10.5555/3305381.3305384.

Parand A Alamdari, Toryn Q Klassen, Elliot Creager, and Sheila A McIlraith. Remembering to be fair: Non-markovian fairness in sequential decision making. In *Forty-first International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/alamdari24a.html.

Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38747–38760. Curran Associates, Inc., 2022. doi: 10.52202/068431-2808. URL https://doi.org/10.52202/068431-2808.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ryan S. Baker and Aaron Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32(4):1052–1092, Dec 2022. ISSN 1560-4306. doi: 10.1007/s40593-021-00285-9. URL https://doi.org/10.1007/s40593-021-00285-9.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019. URL http://jmlr.org/papers/v20/17-612.html.

Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Steven Z Wu. Equal opportunity in online classification with partial feedback. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/084afd913ab1e6ea58b8ca73f6cb41a6-Paper.pdf.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021. doi: 10.1177/0049124118782533.

Trenton Chang and Jenna Wiens. From biased selective labels to pseudo-labels: an expectation-maximization framework for learning from biased decisions. *Proceedings of Machine Learning Research*, 235:6286, 2024. URL https://dl.acm.org/doi/10.5555/3692070.3692313.

Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pp. 181–190. PMLR, 2020. URL http://proceedings.mlr.press/v124/chen20a.html.

Jianfeng Chi, Jian Shen, Xinyi Dai, Weinan Zhang, Yuan Tian, and Han Zhao. Towards return parity in markov decision processes. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1161–1178. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/chi22a.html.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pp. 442–450, Red Hook, NY, USA, 2010. Curran Associates Inc. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/59c33016884a62116be975a9bb8257e3-Paper.pdf.

Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. Causal Modeling for Fairness In Dynamical Systems. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2185–2195. PMLR, July 2020. doi: 10.5555/3524938.3525142. URL https://proceedings.mlr.press/v119/creager20a.html.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 525–534, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372878. URL https://doi.org/10.1145/3351095.3372878.

Zhihong Deng, Jing Jiang, Guodong Long, and Chengqi Zhang. What hides behind unfairness? exploring dynamics fairness in reinforcement learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 3908–3916, 2024. doi: 10.24963/ijcai.2024/432. URL https://doi.org/10.24963/ijcai.2024/432.

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 160–171. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/ensign18a.html.

Jake Fawkes, Nic Fishman, Mel Andrews, and Zachary Chase Lipton. The fragility of fairness: Causal sensitivity analysis for fair machine learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. doi: 10.5555/3737916.3742272. URL https://openreview.net/forum?id=SXYmSTXyHm.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Fair off-policy learning from observational data. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 13943–13972, 2024. URL https://proceedings.mlr.press/v235/frauen24a.html.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022. doi: https://doi.org/10.1111/jofi.13090. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13090.

Usman Gohar, Zeyu Tang, Jialu Wang, Kun Zhang, Peter L Spirtes, Yang Liu, and Lu Cheng. Long-term fairness inquiries and pursuits in machine learning: A survey of notions, methods,

and challenges. arXiv preprint arXiv:2406.06736, 2024. URL https://arxiv.org/abs/2406.06736.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. URL https://dl.acm.org/doi/10.5555/3157382.3157469.

Yaowei Hu and Lu Zhang. Achieving long-term fairness in sequential decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9549–9557, 2022. doi: 10.1609/AAAI.V36I9.21188. URL https://doi.org/10.1609/aaai.v36i9.21188.

Yaowei Hu, Jacob Lear, and Lu Zhang. Striking a balance in fairness for dynamic systems through reinforcement learning. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 662–671. IEEE, 2023. doi: 10.1109/BigData59044.2023.10386299. URL https://doi.org/10.1109/BigData59044.2023.10386299.

INEP. Instituto nacional de estudos e pesquisas educaionais anísio teixeira, microdados do enem, 2025. URL https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acessed in: 2025-09-24.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *International conference on machine learning*, pp. 1617–1626. PMLR, 2017. URL https://proceedings.mlr.press/v70/jabbari17a.html.

Vijay Keswani, Anay Mehrotra, and L. Elisa Celis. Fair classification with partial feedback: an exploration-based data collection approach. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. URL https://proceedings.mlr.press/v235/keswani24a.html.

Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 277–287. PMLR, 2020. URL http://proceedings.mlr.press/v108/kilbertus20a.html.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 275–284, 2017. doi: 10.1145/3097983.3098066. URL https://doi.org/10.1145/3097983.3098066.

Jacob Lear and Lu Zhang. A causal lens for learning long-term fair policies. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rPkCVSsoM4.

Dexun Li and Pradeep Varakantham. Efficient resource allocation with fairness constraints in restless multi-armed bandits. In *Uncertainty in Artificial Intelligence*, pp. 1158–1167. PMLR, 2022. URL https://proceedings.mlr.press/v180/li22e.html.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018. URL http://proceedings.mlr.press/v80/liu18c.html.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative Prediction. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609. PMLR, July 2020. doi: 10.5555/3524938.3525642. URL https://proceedings.mlr.press/v119/perdomo20a.html.

Jansen Silva de Brito Pereira, Giovani Valdrighi, and Marcos Medeiros Raimundo. M^2fgb: A min-max gradient boosting framework for subgroup fairness. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3106–3118, 2025. doi: 10.1145/3715275.3732197. URL https://doi.org/10.1145/3715275.3732197.

ProPublica. How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. (Accessed on 13th Nov. 2025).

Bhagyashree Puranik, Upamanyu Madhow, and Ramtin Pedarsani. A dynamic decision-making framework promoting long-term fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 547–556, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534127. URL https://doi.org/10.1145/3514094.3534127.

Reilly Raab, Ross Boczar, Maryam Fazel, and Yang Liu. Fair Participation via Sequential Policies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14758–14766, March 2024. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v38i13.29394. URL https://ojs.aaai.org/index.php/AAAI/article/view/29394.

Miriam Rateike, Ayan Majumdar, Olga Mineeva, Krishna P Gummadi, and Isabel Valera. Don't throw it away! the utility of unlabeled data in fair decision making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1421–1433, 2022. doi: 10.1145/3531146.3533199. URL https://doi.org/10.1145/3531146.3533199.

Miriam Rateike, Isabel Valera, and Patrick Forré. Designing Long-term Group Fair Policies in Dynamical Systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 20–50, Rio de Janeiro Brazil, June 2024. ACM. ISBN 9798400704505. doi: 10.1145/3630106.3658538. URL https://dl.acm.org/doi/10.1145/3630106.3658538.

US Federal Reserve. Report to the congress on credit scoring and its effects on the availability and affordability of credit, 2007.

Sahand Rezaei-Shoshtari, Hanna Yurchyk, Scott Fujimoto, Doina Precup, and David Meger. Fairness in reinforcement learning with bisimulation metrics. *arXiv preprint arXiv:2412.17123*, 2024. URL https://arxiv.org/abs/2412.17123.

Harsh Satija, Alessandro Lazaric, Matteo Pirotta, and Joelle Pineau. Group fairness in reinforcement learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=JkIH4MeOc3.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Seamus Somerstep, Ya'acov Ritov, and Yuekai Sun. Algorithmic fairness in performative policy learning: Escaping the impossibility of group fairness. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 616–630, 2024. doi: 10.1145/3630106.3658929.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *Advances in Neural Information Processing Systems*, 28, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/39027dfad5138c9ca0c474d71db915c3-Paper.pdf.

Shufan Wang, Guojun Xiong, and Jian Li. Online restless multi-armed bandits with long-term fairness constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15616–15624, 2024. doi: 10.1609/aaai.v38i14.29489. URL https://doi.org/10.1609/aaai.v38i14.29489.

Min Wen, Osbert Bastani, and Ufuk Topcu. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1152. PMLR, 2021. URL https://proceedings.mlr.press/v130/wen21a.html.

Tian Xie and Xueru Zhang. Automating data annotation under strategic human agents: Risks and potential solutions. *Advances in Neural Information Processing Systems*, 37:127436–127482, 2024. doi: 10.52202/079017-4047. URL https://doi.org/10.52202/079017-4047.

Yuancheng Xu, Chenghao Deng, Yanchao Sun, Ruijie Zheng, Xiyao Wang, Jieyu Zhao, and Furong Huang. Adapting static fairness to sequential decision-making: Bias mitigation strategies towards equal long-term benefit rate. In *Forty-first International Conference on Machine Learning*, 2024. URL https://dl.acm.org/doi/10.5555/3692070.3694332.

Tongxin Yin, Reilly Raab, Mingyan Liu, and Yang Liu. Long-term fairness with unknown dynamics. *Advances in Neural Information Processing Systems*, 36:55110–55139, 2023. doi: 10.52202/075280-2407. URL https://doi.org/10.52202/075280-2407.

Eric Yang Yu, Zhizhen Qin, Min Kyung Lee, and Sicun Gao. Policy optimization with advantage regularization for long-term fairness in decision systems. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-0596. URL https://doi.org/10.52202/068431-0596.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020a. URL https://dl.acm.org/doi/10.5555/3495724.3497274.

Yiming Zhang, Quan Vuong, and Keith Ross. First order constrained optimization in policy space. *Advances in Neural Information Processing Systems*, 33:15338–15349, 2020b. URL https://dl.acm.org/doi/10.5555/3495724.3497010.

A NOTATION

Notation	Definition
x	Non-sensitive features
y	Binary label
z	Binary group, sensitive attribute
a	Binary decision, $a = 1$ represents acceptance
$\alpha(x, z)$	Probability of positive label for a sample (x, z)
$\pi(x, z)$	Probability of acceptance for a sample (x, z)
$R(y, a)$	Reward for the decision-maker of decision a for a sample with label y
$U(y, a)$	Reward for the individual with a label y for a decision a
X_t	Random variable for non-sensitive features at time t
Y_t	Random variable for binary label at time t
Z	Random variable for binary group
A_t	Random variable for binary decision at time t
P_0	Starting distribution of features (x, z)
$P_{\mathcal{T}}$	Probability transition function
μ_t^i	Expected utility obtained by group i at time t
$\mathcal{C}^i$	Condition to measure utility of group i
Δ_t	Signed difference of utility between groups at time t
$\Delta_t^{A=1}$	Signed difference of utility between groups at time t conditioned on accepted samples
$\hat{y}$	Predict binary label
$\phi(x, z)$	Probability of predicting positive label for sample (x, z)
$\tilde{y}$	Binary label after data imputation procedure
$\tilde{\mu}_t^i$	Expected utility obtained by group i calculated with imputated labels at time t
$\tilde{\mathcal{C}}^i$	Condition to measure utility of group i
$\tilde{\Delta}_t$	Signed difference of utility between groups after label imputation at time t
ϵ_t^i	Expected signed error of predictions in the rejected population for group i
r_t^i	Rejection rate of group i at time t
$\tilde{\phi}_t^i$	Positive rate of group i after data imputation at time t
$\pi[k]$	Policy deployed at learning iteration k
$\mathrm{w}(x, i)$	Propensity weight for sample (x, i)
$D_R^i(x)$	Probability of observing sample x at the rejected population for group i
$D_A^i(x)$	Probability of observing sample x at the accepted population for group i
$a[1:K]^i$	Acceptance rate of group i over K learning iterations
N^i	Number of samples for accepted for group i
$\hat{\epsilon}_{A,\mathrm{w}}^i$	Inverse propensity weighting estimator of expected signed error of predictions for group i
$d_2(D_R^i \| D_A^i)$	Renyi divergence between rejected and accepted populations for group i
$\bar{\epsilon}^i$	Upper-bound of expected signed error of predictions in the rejected population for group i
$A(s_t, a_t)$	Advantage of executing action a_t in state s_t
$\hat{A}_\beta(s_t, a_t)$	Penalized advantage of executing action a_t in state s_t with weights β
L^{PPO}	Standard PPO loss function
L^{Renyi}	Renyi divergence penalization loss function
$L^{Classif}$	Loss function for classifier ϕ
β_1	Advantage penalization weight
β_2	Renyi penalization weight

Table 1: Summary of notation.

B EXTENDED RELATED WORKS

In this section, we discuss in greater detail related works on long-term fairness and selective labels, and other related areas. For a comprehensive review of long-term fairness, we refer to the survey by Gohar et al. (2024).

Long-term fairness has gained significant attention since the seminal work by Liu et al. (2018), which presented an analysis of fairness policies in a credit scenario with one-step feedback. Following D'Amour et al. (2020) employed simulations to evaluate the effect of fair policies over a larger period. Both works showed that ensuring fairness at each iteration might cause harm in the long term when dynamics are introduced.

Algorithmic solutions commonly leverage reinforcement learning solutions or causal modeling. Considering that feedback dynamics are known, Wen et al. (2021) introduced fairness metrics to the MDP setting by formulating individuals' rewards as a second objective, and Rateike et al. (2024) studied settings where a fixed-threshold policy can converge to a fair equilibrium. Works have also formulated fairness as a time-dependent cost, which is aggregated over time with discounts, defining value functions of unfairness (Satija et al., 2023; Xu et al., 2024). Particularly, Xu et al. (2024) introduced the idea of defining group-wise supply and demands, and evaluating fairness based on the ratio of such group-wise measures. Jabbari et al. (2017) analyzes a meritocratic notion of fairness in RL where an action cannot be preferred unless its long-term return is higher, highlighting the difficulty of obtaining reliable long-term value estimates. This configuration of long-term disparity also permits the use of standard algorithms for constrained MDPs (CMDPs), such as CPO (Achiam et al., 2017) and FOCOPS (Zhang et al., 2020b), which perform projections of the policy to the feasible set at each learning iteration. A set of works has studied the PPO algorithm to ensure fairness. Yu et al. (2022) and Hu et al. (2023) included a penalization term on the advantage estimate used for policy optimization, while Lear & Zhang (2025) used an expansion of the disparity in qualification as a value function. Q-learning was adapted for long-term fairness by Chi et al. (2022) and Alamdari et al. (2024).

The relation between short-term fairness and long-term fairness has also been studied by previous works (Hu et al., 2023; Alamdari et al., 2024; Lear & Zhang, 2025). Yin et al. (2023) used a different framework where states were the joint distribution of the population. To support continuous state and actions, it employed a modification of the least-squares value iteration algorithm. Deng et al. (2024) proposed learning an approximation of dynamics to evaluate if unfairness is introduced by them, that is, if a fair state can transition to an unfair state even if decisions are group-independent. A subset of works for long-term fairness considered different dynamics between decisions and population distributions, where the participation of groups was not fixed over time and depended on the quality of predictions (accuracy) or on the acceptance rates (Puranik et al., 2022; Raab et al., 2024). All of these approaches considered only measuring fairness from fully observable features X (no use of labels Y).

In the strategic classification setting, the deployment of a decision model induces the update of data distribution by the adaptation of individuals to the decision rationale. With these dynamics, Xie & Zhang (2024) studied how the iterative update of the model with strategic data could result in long-term disparity, even when fairness is enforced at each iteration. Somerstep et al. (2024) studied how such dynamics could be used to shift the data distribution to a fair state. However, works in strategic classification consider a single timestep of dynamics.

In the stochastic K-out-of-N bandit model, the decision-maker at each iteration must select K arms over N total possibilities and observes rewards only for those arms. Long-term fairness has already been discussed in this setting by considering that each arm belongs to a group, and that each group should be selected (any arm of the group) with a frequency higher than a threshold Chen et al. (2020); Li & Varakantham (2022); Wang et al. (2024). While these works handle partial feedback, the classical bandit assumption that actions do not influence future contexts eliminates long-term feedback loops that motivate our work.

Partial-label scenarios have been analyzed in simpler decision-theoretic models or in settings with time-invariant data distributions. Zhang et al. (2020a) presented a theoretical study of threshold policies that satisfy fairness in the short-term, but not necessarily in the long-term. While a partial observation MDP was used in the analysis, it did not consider learning in such a setting. Fawkes et al. (2024) audit benchmark fairness datasets and reported that selection bias (a class of bias that includes partial feedback) was identified in 85% of them. In static environments, previous works considered the problem of sequentially employing a policy that is used to learn the unseen data distribution with selective labels. Kilbertus et al. (2020) showed policies should be able to "explore" so that a learning algorithm does not end in a suboptimal utility and fairness. Following Rateike et al. (2022) considered using the unlabeled data to learn an unbiased representation of individuals'

features, which were then used to train the policy. Keswani et al. (2024) presented an algorithm to learn the optimal policy with suboptimal estimates of labels. Chang & Wiens (2024) introduced a similar approach in the time-invariant setting, where a model is used to learn the treatment policy and another to learn the outcome. Their solution is an expectation-maximization algorithm that iteratively trains both models. In supervised learning, Lakkaraju et al. (2017) study selective labels in classification when data is collected from heterogeneous human decision-makers with different acceptance rates, and propose using data from lenient decision-makers to estimate failure rates of a new black-box model; however, their analysis does not consider that data is dynamically affected by decisions. Frauen et al. (2024) consider fairness in off-policy learning, aiming to learn a fair policy using data collected by a possibly discriminatory behavior policy; they assume features are static over time and learn a representation independent of the sensitive attribute while preserving predictive information, leveraging IPW similarly to our approach.

Causal modeling provides a language for defining the feedback loops that induce long-term disparity. Creager et al. (2020) discussed the benefits of representing assumptions within the causal diagrams' framework, providing various examples where an undesired effect occurs when the causal structure of the system is misspecified. One such analysis was of off-policy evaluation in the setting of partial feedback, yet their work did not include a theoretical analysis. Hu & Zhang (2022) connected causality and performative predictions in long-term fairness by transforming an optimization problem defined by a causal model to a problem of performative prediction.

C PROOFS

In this section, we will omit the subscript t whenever it is not relevant. Furthermore, we simplify the notation $P^i(E|C) := P(E|C, Z = i)$ for any event E and condition C. We will also write $\mathbb{E}^i[E|C] := \mathbb{E}^i[E|C, Z = i]$.

C.1 PROOF OF PROPOSITION 3.1

We first write the proposition presented using a formal notation.

Proposition C.1 (Restatement of Proposition 3.1)**.** *Let $a_t^i = P(A_t = 1|Z = i)$ be the acceptance rate of group i. For each fairness principle, the disparity calculated from the accepted population $\Delta_t^{A=1}$ has the decomposition $\Delta_t^{A=1} = (\mu_t^1 c^1 - \mu_t^0 c^0) + (d^1 - d^0)$ where the terms c^i, d^i are:*

- *Qualification parity: $c^i = P(A_t = 1|Y_t = 1, Z = i)/a_t^i, \;\; d^i = 0$.*
- *Accuracy parity: $c^i = 1/a_t^i, \;\; d^i = -P^1(Y_t = 0, A_t = 0|Z = i)/a_t^i$.*
- *Equality of opportunity: $c^i = d^i = 0$ (that is, $\Delta_t^{A=1} = 0$ always).*

And $|\Delta_t^{A=1}| = 0$ is not a sufficient condition for $|\Delta_t| = 0$.

Proof. First, we consider each fairness principle and identify an expression for $\Delta_t^{A=1}$:

1) Qualification parity

By considering each term of $\Delta^{A=1}$:

$$\mu^i = P^i(Y = 1|A = 1) = P^i(Y = 1)\frac{P^i(A = 1|Y = 1)}{a^i} \tag{8}$$

And by joining both terms, we have:

$$\Delta^{A=1} = P^1(Y = 1|A = 1) - P^1(Y = 1|A = 1) \tag{9}$$

$$= P^1(Y = 1)\frac{P^1(A = 1|Y = 1)}{a^1} - P^0(Y = 1)\frac{P^0(A = 1|Y = 1)}{a^0} \tag{10}$$

2) Accuracy parity

Similarly, considering each side $\Delta^{A=1}$:

$$P^i(Y = A) = P^i(Y = 1, A = 1) + P^i(Y = 0, A = 0) \tag{11}$$
$$= P^i(Y = A|A = 1)a^i + P^i(Y = 0, A = 0) \implies \tag{12}$$
$$P^i(Y = A|A = 1) = \frac{P^i(Y = A)}{a^i} - \frac{P^i(Y = 0, A = 0)}{a^i} \tag{13}$$

And by joining both terms:

$$\Delta^{A=1} = P^1(A = Y|A = 1) - P^0(A = Y|A = 1) \tag{14}$$
$$= \mu^1/a^1 - \mu^0/a^0 - \left(\frac{P^1(Y = 0, A = 0)}{a^1} - \frac{P^0(Y = 0, A = 0)}{a^0}\right) \tag{15}$$

3) Equality of opportunity

It is straightforward to see that $P^i(A = 1|Y = 1 \wedge A = 1) = 1$, concluding that $\Delta^{A=1} = 1 - 1 = 0$ independently of the real disparity Δ.

Conclusion

Now, if $|\Delta_t^{A=1}| = 0$ we can have $|\Delta_t| > 0$ by setting:

- Qualification parity: $c^1 \neq c^0$ and $\mu_t^1 = (c^1/c^0)\mu_t^0$ which implies $\mu^1 \neq \mu^0 \implies |\Delta_t| > 0$.
- Accuracy parity: $d^1 = d^0, c^1 \neq c^0$ and $\mu_t^1 = (c^1/c^0)\mu_t^0$ which implies $\mu^1 \neq \mu^0 \implies |\Delta_t| > 0$.
- Equality of opportunity is direct, as $\Delta_t^{A=1} = 0$ always.

□

C.2 Proof of Theorem 3.1

Proof. We will define the random variable $\epsilon = \hat{Y} - Y$, $\epsilon \in \{-1, 0, 1\}$ and use the relation $\tilde{Y} = Y + (1 - A)\epsilon$. Based on this, we can conclude:

$$\mathbb{E}^i[(1 - A)\epsilon] = \underbrace{\mathbb{E}^i[(1 - A)\epsilon \mid A = 1]}_{=0} a^i + \mathbb{E}^i[(1 - A)\epsilon \mid A = 0]r^i \tag{16}$$
$$= \mathbb{E}^i[\epsilon \mid A = 0]r^i = \epsilon^i r^i \tag{17}$$

With ϵ^i as defined in the section. Then, we consider each fairness principle.

1) Equality of qualification

Considering each term of $\tilde{\Delta}$, we have that:

$$\mathbb{E}^i[\tilde{Y}] = \mathbb{E}^i[Y + (1 - A)\epsilon] = \mathbb{E}^i[Y] + \mathbb{E}^i[(1 - A)\epsilon] \tag{18}$$

We combined both terms to rewrite $\tilde{\Delta}$:

$$\tilde{\Delta} = (\mathbb{E}^1[Y] + \epsilon^1 r^1) - (\mathbb{E}^0[Y] + \epsilon^0 r^0) \tag{19}$$
$$= (\mathbb{E}^1[Y] - \mathbb{E}^0[Y]) + (\epsilon^1 r^1 - \epsilon^0 r^0) = \Delta + (\epsilon^1 r^1 - \epsilon^0 r^0) \tag{20}$$

2) Equality of accuracy

Considering each term of $\tilde{\Delta}$, we have that:

$$\begin{aligned}
&\mathbb{E}^i[1\{A = \tilde{Y}\}] = && (21)\\
&= P^i(A = 1, \tilde{Y} = 1) + P^i(A = 0, \tilde{Y} = 0) && (22)\\
&= P^i(A = 1, Y = 1) + P^i(A = 0, Y + \epsilon = 0) && (23)\\
& && (24)
\end{aligned}$$

Let's work on the term $P^i(A = 0, Y + \epsilon = 0)$:

$$\begin{aligned}
&P^i(A = 0, Y + \epsilon = 0) = r^i P(Y + \epsilon = 0 \mid A = 0) && (25)\\
&= r^i \mathbb{E}^i[1 - (Y + \epsilon) \mid A = 0] && (26)\\
&= r^i(1 - \mathbb{E}^i[Y \mid A = 0] - \epsilon^i) && (27)\\
&= r^i - \mathbb{E}^i[Y \mid A = 0]r^i - \epsilon^i r^i && (28)\\
&= \underbrace{r^i - P^i(Y = 1, A = 0)}_{P^i(Y=0,A=0)} - \epsilon^i r^i && (29)\\
&= P^i(Y = 0, A = 0) - r^i \epsilon^i && (30)
\end{aligned}$$

Replacing it in $\mathbb{E}^i[1\{A = \tilde{Y}\}]$:

$$\begin{aligned}
\mathbb{E}[1\{A = \tilde{Y}\} \mid Z = z^\bullet] &= P^i(A = 1, Y = 1) + P^i(Y = 0, A = 0) - r^i \epsilon^i && (31)\\
&= \mathbb{E}[1\{A = Y\} \mid Z = z^\bullet] - r^i \epsilon^i && (32)
\end{aligned}$$

Then, we have that by replacing both terms of $\tilde{\Delta}$.

$$\begin{aligned}
\tilde{\Delta} &= \mathbb{E}^1[1\{A = \tilde{Y}\}] - \mathbb{E}^0[1\{A = \tilde{Y}\}] = && (33)\\
&= \left(\mathbb{E}^1[1\{A = Y\}] - r^1\epsilon^1\right) - \left(\mathbb{E}^0[1\{A = Y\}] - r^0\epsilon^0\right) = && (34)\\
&= \Delta - (\epsilon^1 r^1 - \epsilon^0 r^0) && (35)
\end{aligned}$$

3) Equality of opportunity

We first open one term of $\tilde{\Delta}$:

$$\mathbb{E}^i[A = 1 \mid \tilde{Y} = 1] = P^i(A = 1 \mid \tilde{Y} = 1) = \frac{P^i(A = 1, \tilde{Y} = 1)}{P^i(\tilde{Y} = 1)} \tag{36}$$

Notice that $P^i(A = 1, \tilde{Y} = 1) = P^i(A = 1, Y = 1)$ as $\tilde{Y} = Y$ when the action is positive. We are now interested in replacing the denominator $P^i(\tilde{Y} = 1)$ to $P^i(Y = 1)$. To do so, we can define $\kappa^i = \dfrac{P^i(Y = 1)}{P^i(\tilde{Y} = 1)}$ with the assumption that $P^i(\tilde{Y} = 1) \neq 0$ and obtain:

$$\mathbb{E}^i[A = 1 \mid \tilde{Y} = 1] = \frac{P^i(A = 1, Y = 1)}{P^i(Y = 1)}\kappa^i = \mathbb{E}^i[A = 1 \mid Y = 1]\kappa^i \tag{37}$$

This shows that the true positive rate calculated from the observed labels is equal to the true positive rate with the multiplying factor κ^i, that is, the ratio of real positive labels and observed positive labels. Then, joining both terms in the expression of $\tilde{\Delta}$, we obtain:

$$\tilde{\Delta} = \mathbb{E}^1[A = 1 \mid Y = 1]\kappa^1 - \mathbb{E}^0[A = 1 \mid Y = 1]\kappa^0 \tag{38}$$

We are also interested in rewriting κ^i to remove the direct dependence on Y, a value that is partially observed. We have that:

$$\begin{aligned}
\mathbb{E}^i[\tilde{Y}] =& \mathbb{E}^i[Y] + \mathbb{E}^i[(1-A)\epsilon] && (39)\\
\implies P^i(Y=1) =& \mathbb{E}^i[\tilde{Y}] - \mathbb{E}^i[(1-A)\epsilon] && (40)\\
=& P^i(\tilde{Y}=1) - \epsilon^i r^i && (41)
\end{aligned}$$

And then:

$$\kappa^i = \frac{P^i(\tilde{Y}=1) - \epsilon^i r^i}{P^i(\tilde{Y}=1)} = 1 - \frac{\epsilon^i r^i}{\tilde{\phi}^i} \tag{42}$$

With $\tilde{\phi}^i$ defined as in the section.

□

C.3 Proof of Theorem 3.2

Proof. We first consider the scenario of qualification parity and accuracy parity. From Theorem 3.1, we have that:

$$\begin{aligned}
\tilde{\Delta} &= \Delta \pm (r^1\epsilon^1 - r^0\epsilon^0) \implies && (43)\\
|\Delta| &= |\tilde{\Delta} \pm (r^1\epsilon^1 - r^0\epsilon^0)| && (44)\\
&\leq |\tilde{\Delta}| + |r^1\epsilon^1 - r^0\epsilon^0| && (45)\\
&\leq \omega/2 + \omega/2 = \omega && (46)
\end{aligned}$$

Where the first two lines use $\pm$ due to the different expressions obtained for qualification parity and accuracy parity.

Now, with the equality of opportunity fairness principle, we have from Theorem 3.1:

$$\begin{aligned}
\tilde{\Delta} &= \mu^1\kappa^1 - \mu^0\kappa^0 && (47)\\
&= \kappa^1\Delta + \mu^0(\kappa^1 - \kappa^0) \implies && (48)\\
\kappa^1|\Delta| &= |\tilde{\Delta} - \mu^0((1 - r^1\epsilon^1/\tilde{\phi}^1) - (1 - r^0\epsilon^0/\tilde{\phi}^0))| && (49)\\
&= |\tilde{\Delta} - \mu^0(-r^1\epsilon^1/\tilde{\phi}^1 + r^0\epsilon^0/\tilde{\phi}^0)| && (50)\\
&\leq |\tilde{\Delta}| + \mu^0|r^1\epsilon^1/\tilde{\phi}^1 - r^0\epsilon^0/\tilde{\phi}^0| && (51)\\
&\leq |\tilde{\Delta}| + |r^1\epsilon^1/\tilde{\phi}^1 - r^0\epsilon^0/\tilde{\phi}^0| && (52)\\
&\leq \frac{(1-v)\omega}{2} + \frac{(1-v)\omega}{2} = (1-v)\omega \implies && (53)\\
|\Delta| &\leq \frac{(1-v)\omega}{\kappa^1} \leq \frac{(1-v)\omega}{1-v} = \omega && (54)
\end{aligned}$$

Where line 52 uses the fact that $\mu^0 \leq 1$.

□

C.4 PROOF OF THEOREM 3.3

Theorem 3.3 was initially presented by Cortes et al. (2010). Here we present the original statement and describe the adaptation to our scenario. First we define models h, which are evaluated from a bounded loss $L(h(x), f(x))$ (abbreviated by $L_h(x)$), the risk $R(h) = \mathbb{E}_{x\sim P}[L_h(x)]$ and the weighted empirical loss $\hat{R}_{\mathrm{w}}(h) = \sum_{i=1}^{m} \mathrm{w}(x_i)L_h(x^i)$ calculated from m i.i.d. samples (x_i, y_i) obtained by distribution Q.

Theorem C.1 (Theorem 3 from Cortes et al. (2010))**.** *Let H be a hypothesis set such that Pdim($\{L_h(x) : h \in H\}$) = $p < \infty$. Assume that $d_2(P||Q) < \infty$ and $\mathrm{w}(x) = P(x)/Q(x) \neq 0$ for all x. Then, for any $\delta > 0$, with probability of at least $1 - \delta$, the following holds:*

$$R(h) \leq \hat{R}_{\mathrm{w}}(h) + 2^{5/4}\sqrt{d_2(P||Q)}\sqrt[\frac{3}{8}]{\frac{p\log\frac{2me}{p} + \log\frac{4}{\delta}}{m}} \tag{55}$$

In our setting, we evaluated the models ϕ using data collected from previously accepted individuals, that is, $Q := D_A^i$ and:

$$D_A^i(x) = \frac{\left(1 - \prod_{k=1}^{K}(1 - \pi[k](x,i))\right) g(x,i)}{a[1:K]^i} \tag{56}$$

where $g(x,i) := P(X = x|Z = i)$. However, we wish to know the error from the distribution of rejected individuals, which is $P := D_R^i$:

$$D_R^i(x) = \frac{(1 - \pi[K](x,i))g(x,i)}{r^i} \tag{57}$$

and $\mathrm{w}(x,i) = P(X)/Q(x) = D_R^i(x)/D_A^i(x)$ has the expression presented in Sec. 3. Lastly, our loss measure is $\epsilon(x,i) = \mathbb{E}[\hat{Y} - Y|X = x, Z = i]$, which is also bounded, but has support in $[-1, 1]$. With this configuration, $R(h) := \epsilon^i$ and $\hat{R}_{\mathrm{w}}(h) = \hat{\epsilon}^i_{A,\mathrm{w}}$. While the larger support changes the formulation of the bound in Eq. 55, only the constants are different, and big-O is kept the same.

C.5 PROOF OF THEOREM3.4

Proof. Initially, as $\overline{\epsilon}^i \geq \epsilon^i$ and $r^i > 0, \forall i$, we have that $\sum_i r^i|\epsilon^i| \leq \sum_i r^i|\overline{\epsilon}^i|$. By leveraging results from Theorem 3.1, we have that for qualification parity and accuracy parity:

$$\tilde{\Delta} = \Delta \pm (r^1\epsilon^1 - r^0\epsilon^0) \implies \tag{58}$$
$$|\Delta| = |\tilde{\Delta} \pm (r^1\epsilon^1 - r^0\epsilon^0)| \tag{59}$$
$$\leq |\tilde{\Delta}| + |r^1\epsilon^1 - r^0\epsilon^0| \tag{60}$$
$$\leq |\tilde{\Delta}| + |r^1\epsilon^1| + |r^0\epsilon^0| \tag{61}$$
$$\leq |\tilde{\Delta}| + |r^1\overline{\epsilon}^1| + |r^0\overline{\epsilon}^0| \tag{62}$$
$$\leq \omega/2 + 2\omega/4 = \omega \tag{63}$$

And for equality of opportunity:

$$\tilde{\Delta} = \mu^1\kappa^1 - \mu^0\kappa^0 = \kappa^1\Delta + \mu^0(\kappa^1 - \kappa^0) \implies \tag{64}$$

$$\kappa^1|\Delta| = |\tilde{\Delta} - \mu^0(\kappa^1 - \kappa^0)| \tag{65}$$

$$\leq |\tilde{\Delta}| + \mu^0|\kappa^1 - \kappa^0| \tag{66}$$

$$\leq |\tilde{\Delta}| + |\kappa^1 - \kappa^0| = |\tilde{\Delta}| + \left|r^1\epsilon^1/\tilde{\phi}^1 - r^0\epsilon^0/\tilde{\phi}^0\right| \tag{67}$$

$$\leq |\tilde{\Delta}| + \sum_i r^i\epsilon^i/\tilde{\phi}^i \tag{68}$$

$$\leq |\tilde{\Delta}| + \sum_i r^i\bar{\epsilon}^i/\tilde{\phi}^i \tag{69}$$

$$\leq \frac{(1-v)\omega}{2} + \frac{(1-v)\omega}{2} = (1-v)\omega \tag{70}$$

$$|\Delta| \leq \frac{(1-v)\omega}{\kappa^1} \leq \frac{(1-v)\omega}{(1-v)} = \omega \tag{71}$$

□

D ALGORITHM

Algorithm 1: SELLF

Initialize neural networks π, ϕ, V with respective weights $\theta_\pi^0, \theta_\phi^0, \theta_V^0$ and memory buffer $M = \{\}$;
for $k = 1, 2, \ldots, K$ **do**
  Initialize replay buffer $B = \{\}$;
  for *episode* = $1, \ldots, E$ **do**
    for $t = 1, 2, \ldots, T$ **do**
      Sample $a^t \sim \pi(x^t, z)$, $y^t \sim \alpha(x^t, z)$, $x^{t+1} \sim P_{\mathcal{T}}(x^t, z, a^t, y^t)$, $\hat{y}^t \sim \phi(x^t, z)$
      Run data imputation $\tilde{y}^t \leftarrow a^t y^t + (1 - a^t)\hat{y}^t$
      $B \leftarrow B \cup \{z, x^t, \tilde{y}^t, a^t, r^t, x^{t+1}, \tilde{\Delta}^t\}$
      $M \leftarrow M \cup \{x^t, z, \tilde{y}^t\}$ if $a^t = 1$
    end
  end
  for *each predictor gradient step* **do**
    Sample mini-batch from M
    $\theta_\phi^k \leftarrow \theta_\phi^k - \gamma\nabla_{\theta_\phi}\sum_{i\in\{0,1\}}\hat{\mathbb{E}}_{D_A^i}[\mathrm{w}(x_t, i)\ell(\tilde{y}_t, \phi(x_t, i))/\mathrm{w}(i)]$
  end
  for *each policy gradient step* **do**
    $\hat{A}_\beta(s_t, a_t) \leftarrow \hat{A}(s_t, a_t) - \beta_1 \max\{|\tilde{\Delta}_t| - \omega/2, 0\}$
    $d(\theta_\pi) \leftarrow \pi(x^t, z)/\pi_{\theta_\pi^t}(x^t, z)$
    $J^{\text{CLIP}}(\theta_\pi) \leftarrow \hat{\mathbb{E}}[\min(d(\theta_\pi)\hat{A}(s_t, a_t), \text{clip}(d(\theta_\pi), 1-\epsilon, 1+\epsilon)\hat{A}_\beta)]$
    $L^{Renyi}(\theta_\pi) \leftarrow (r_t^1\hat{\mathbb{E}}[\mathrm{w}_t^2|Z=1] + r_t^0\hat{\mathbb{E}}[\mathrm{w}_t^2|Z=0])/2$
    $\theta_\pi^{t+1} \leftarrow \theta_\pi^t + \gamma(\nabla_{\theta_\pi}J^{\text{CLIP}}(\theta_\pi) - \nabla_{\theta_\pi}L^{Renyi}(\theta_\pi))$
    $\theta_V^{t+1} \leftarrow \theta_V^t - \alpha\nabla_{\theta_V}\mathbb{E}[(V(s^t) - G(s^t))^2]$; ▷ $G(s^t) \leftarrow \sum_{i=0}^T \gamma^i r_{t+i}$
  end
end

E DATASETS AND ENVIRONMENTS

This work considers the effects of algorithms on the distribution of population attributes. This characteristic impedes the evaluation of algorithms in historical (and static) data, as they will not present the effects of the intervention of algorithms. For that reason, we employ semisynthetic

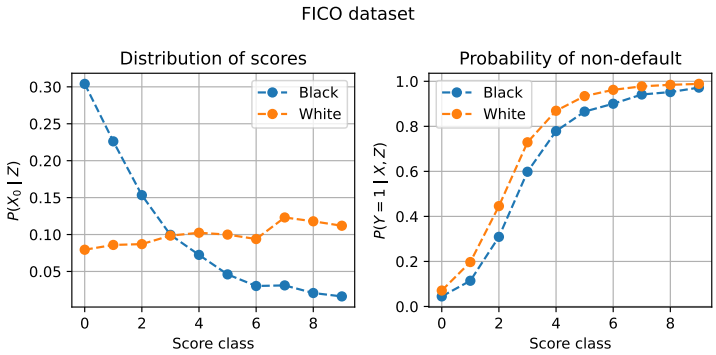


Figure 5: Probability distributions calculated from the FICO dataset to define the environment.

environments to evaluate the proposed algorithms, which is commonly done in studies of long-term fairness. The environments have initial distributions of variables (X, Z) and the relation with the target Y calculated from real-world datasets. To model the environment's dynamics, we assume transition functions based on features, labels, and decisions. These dynamics must be plausible for the system modeled, of which we considered three: loan applications based on FICO, crime recidivism based on COMPAS, and school admission based on ENEM.

Lending FICO (Reserve, 2007) is a common open-source dataset utilized in fairness studies. It consists of anonymized profiles of clients of a banking institution with a credit score that was calculated from these attributes. Using the data available from Barocas et al. (2023), race was defined as the sensitive attribute Z, using two classes ("Black" and "white"). For simplicity, we set each group with probability 0.5. Then, for each group, we calculate the probability of observing each score (from 10 possible discretized score values). This was then used as $P(X^0|Z)$. Next, we calculate the probability of payment given each score for each group, that is $\alpha(X, Z) = P(Y = 1|Z, X)$. Both distributions are presented in Fig. 5. It is possible to see that while the white population is almost uniformly spread among scores, almost 50% of the Black individuals have a score class of 0 or 1. When considering the probability of payment, we can see that both groups present very similar behavior, yet a small difference is present. We observe that the probability of payment of a Black individual of the same score class as a white one is smaller. This might be caused by external social aspects that were not fully captured by the credit score.

In the FICO environment, we used the dynamics first presented by Liu et al. (2018). If an individual is rejected, the credit score is kept the same. If an individual is accepted, their credit score will increase by one unit if $Y = 1$ and decrease by one unit if $Y = -1$.

Criminal Recidivism COMPAS (ProPublica) is a software used in courts in the US to assess the likelihood of recidivism. A study of great importance by ProPublica showed that this tool consistently predicted a higher likelihood for African-Americans, indicating a discriminatory practice. In this environment, we leveraged the dataset published by ProPublica, and followed the construction of the environment as done by Zhang et al. (2020a); Rateike et al. (2024). We define two sensitive groups based on race, which are African-American and Caucasian, and use only two variables: age (discretized in 5 classes) and the number of priors (discretized in 8 classes). This results in a setting where X, after the one-hot encoding of variables, has a dimension of 13. The decision-maker must decide between jail $(A = 0)$ and bail $(A = 1)$, and is negatively rewarded if bailed individuals reoffend $(Y = 0)$. We calculate the probability of recidivism from the dataset, which is depicted in Fig. 6. The distribution of $P(X|Z)$ was also obtained from the dataset. We consider a simplified dynamic, where individuals only have their features X altered in the case of bail decision and recidivism, in which we move the priors count to the following group.

School Admission The ENEM is a national exam applied in Brazil that serves as a score for public universities' admission process. Yearly, the data collected from applicants is shared with suitable anonymization procedures. This dataset has recently been used in fairness studies (Pereira et al., 2025; Alghamdi et al., 2022). We use this dataset to model a decision-making process where a

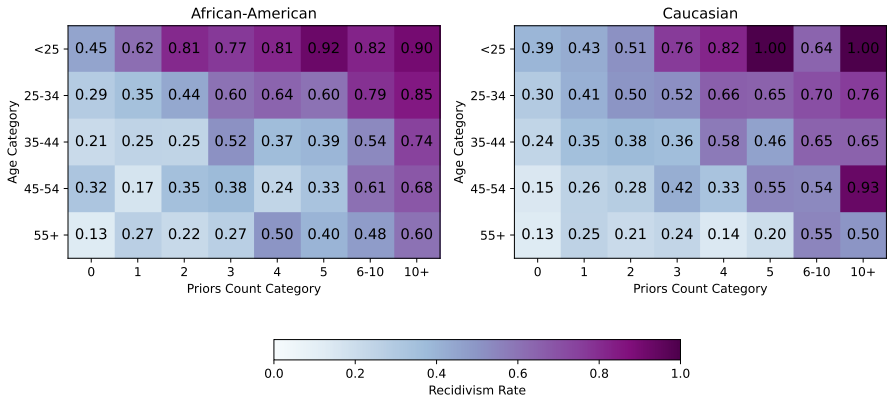


Figure 6: Probability distributions calculated from the COMPAS dataset to define the environment.

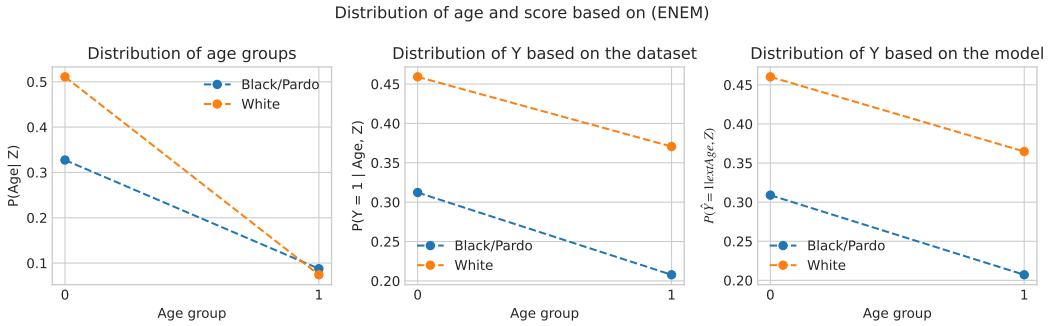


Figure 7: Estimated distributions from ENEM dataset.

decision-maker must accept/reject applicants based on attributes X for a preparatory program. The label Y represents reaching a grade higher than a threshold on the exam, which is only known for individuals who were participants in the program. The decision-maker has the cost of 0.4 for an accept, and is rewarded by accepting applicants with $Y_t = 1$ (applications with Y_t increase the reputation of the preparatory program). We use socioeconomic indicators as the attributes X. Using a random sample of 10,000 applicants from the state of São Paulo, we define $Y = 1$ if the score is higher than 575 and 0 otherwise, which resulted in a probability of positive label of 37%. The sensitive attribute Z is defined as the race attribute with two classes ("white" and "Black/brown") with 62% and 38% of occurrence, respectively. X is composed of 38 categorical features which are one-hot encoded to a 126 dimensional vector.

The dynamics of this environment are defined to simulate the effect of age and of the preparatory program on Y. X contains multiple features, one of them being a categorical age attribute with two categories (see Fig. 7 for the distribution of age categories). We consider that each iteration, the age of the applicant will increase (and other features will be kept the same), and this will affect their qualification, as displayed in the figure. We also add an extra indicator feature on the individual, that will be 1 if they have already been previously accepted or if they have previously had the label $Y = 1$, which increases the probability of the positive label by 0.5 on the following iterations. To simulate $\alpha(X, Z)$, we fit a logistic regression from features X, Z to the label Y. Then, whenever we updated a feature of a candidate, we used an inference with the logistic regression to obtain the probability of them having a positive label. In Fig. 7 we display the distribution of qualification among the two groups, the average qualification for each age category, and the average predicted qualification learned with the logistic model.

F EXPERIMENTAL SETTING

Implementation details All algorithms and experiments were implemented using Python and PyTorch. We followed the implementation of PPO from Stable Baselines 3[4]. The environment implementation follows D'Amour et al. (2020) and is based on Gym[5]. The algorithm POCAR was also used from the original implementation by Yu et al. (2022). The learning hyperparameters for all algorithms were as follows:

- Number of steps in data collection: 2048.
- Mini-batch size: 64.
- Epochs of policy update: 10.
- Gradient steps of predictor after data collection: 25.
- Learning rate: 10^{-5} for policy network and 10^{-2} for predictor network with exponential decay of 0.95.
- Policy π architecture: linear layer $dim(X, Z) \times 64$, Tanh activation, linear layer 64×64, Tanh activation, linear layer 64×1. The value network has the same architecture.
- Predictor ϕ architecture: Linear layer $dim(X, Z) \times 1$. Sigmoid activation is used in the output to obtain probabilities.

To ensure computation efficiency, we randomly select 10 previous policies $\pi[k]$ to calculate the probability of acceptance at previous iterations.

Each algorithm was trained only once and evaluated in the environment with 10 different random seeds. Results are an average of the 10 repetitions.

POCAR Algorithm Yu et al. (2022) proposed advantage regularization for fairness, considering the unfairness of each (state, action) pair and the decrease of unfairness over transitions, using the following expression:

$$\hat{A}_\beta(s_t, a_t) = \hat{A}(s_t, a_t) - \beta_1 \max\{|\Delta_t| - \omega, 0\} - \beta_2 \begin{cases} \max\{|\Delta_{t+1}| - |\Delta_t|, 0\} \text{ if } |\Delta_t| > \omega \\ 0 \text{ otherwise} \end{cases} \tag{72}$$

The first term is similar to the approach used in SELLF, but it also includes a secondary term with weight β_2 that is activated whenever the disparity $|\Delta_t|$ is higher than the threshold ω. This secondary term penalizes the advantage whenever the action increases the disparity from t to $t+1$. This second term could also be incorporated in SELLF, but we opted to remove it for simplicity.

ELBERT Xu et al. (2024) formalized long-term fairness with group-wise supply S_g and demand D_g variables, which can be used to define common fairness metrics. For example, accuracy parity can be defined by using demand as the total number of individuals in the group and supply as the number of correct predictions between individuals for the same group. Then, the disparity measure is equal to $\Delta = S_1/D_1 - S_0/D_0$. To include the long-term effect, it also defines value functions of supply and demand, which are used to calculate a penalized advantage value on the proposed ELBERT algorithm. As ELBERT is unaware of selection bias, we altered supply and demand variables to be calculated only within the selected population.

FOCOPS Proposed by Zhang et al. (2020b), FOCOPS is a constrained policy optimization algorithm inspired by TRPO. For each state, there is an associated cost (in our scenario, disparity), and a policy should satisfy constraints on the long-term discounted cost. During learning iterations, it searches for the optimal policy that satisfies constraints by projecting such a policy in the parameter space. In practice, they leverage Lagrangian weights to enforce constraints, which increase over time as constraints are violated. Similar to other baselines, FOCOPS does not consider the partial observation of labels. For that reason, we used $\tilde{\Delta}_t$ as the cost measure.

[4] https://stable-baselines3.readthedocs.io/en/master/

[5] https://github.com/openai/gym

Algorithm 2: Hyperparameter selection

$L_\Delta \leftarrow$ list of values $|\Delta|^{clip}$ for each hyperparameter configuration;
$L_R \leftarrow$ list of values R_T for each hyperparameter configuration;
$L \leftarrow [\ ]$;
for $|\Delta|^{clip}, R_T$ *in* L_Δ, L_R **do**
  if $|\Delta|^{clip} = \min L_\Delta$ **then**
    L.append(R_T)
  else
    L.append(0)
  end
  return *Hyperparameter configuration with highest value in* L
end

Hyperparameters Optimization For POCAR and SELLF, we evaluated 12 different combinations of values of β_1, β_2. In both algorithms, β_1 sets the weight of the penalization of the disparity measure in the advantage, and was evaluated in $\{1, 2, 5, 10\}$. For POCAR, β_2 was evaluated in $\{1, 2, 5\}$ and for SELLF $\beta_2 \in \{0.01, 0.05, 0.1\}$. SELLF (Semi-sto.) was tuned with the same values of hyperparameters. ELBERT also includes a weight β in the advantage penalization, in which we performed hyperparameter optimization on values $\beta \in \{1, 5, 10, 200, 2000, 20000, 200000\}$, following experiments from the original paper. FOCOPS was evaluated using the ν_{max} (the maximum value of the dual variable) in $\{2, 5, 20, 100\}$ and the long-term constraint in $\{5, 10, 15\}$. Hyperparameters of POCAR with and without oracle were tuned separately.

The selected hyperparameter configuration was the one with the highest reward that reached disparity below ω (0.05) or, if no solution reached such disparity, the one that had minimal disparity. To avoid contamination, algorithms were not given access to the true disparity measure; that is, PPO and POCAR had their hyperparameters tuned based on $\Delta^{A=1}$, POCAR (Oracle) with Δ, and SELLF with $\tilde{\Delta}$. In more detail, we set $|\Delta| = \frac{1}{T}\sum_{i=1}^{T} |\Delta_t|$ (with the respective variation of the disparity measure) which was clipped $|\Delta|^{clip} = \min\{\Delta - \omega, 0\}$. Then, for each algorithm, hyperparameters were selected following Alg. 2

At Tab. 2 is displayed the optimal hyperparameter configuration for SELLF obtained for all environments (including extra configurations from Appendix H.4). Values of $\beta_1 \in \{1, 2\}$ were the optimal ones in 8 of the 9 cases and $\beta_2 = 0.05$ was the optimal value in 5 of the 9 cases, being indicated as default configurations for SELLF.

Environment	Utility types	β_1	β_2
Lending	Accuracy	1	0.05
Lending	Qualification	1	0.1
Lending	Equality of Opportunity	5	0.01
Crime recidivism	Accuracy	2	0.05
Crime recidivism	Qualification	2	0.05
Crime recidivism	Equality of Opportunity	1	0.01
School admission	Accuracy	1	0.05
School admission	Qualification	2	0.1
School admission	Equality of Opportunity	1	0.05

Table 2: Parameter configurations for different experimental environments.

G TABLE OF EXPERIMENTS FROM SEC. 5

Tab. 3 displays the average disparity and accumulated reward for agents. Additional results with varying fairness notions are presented in Appendix H. In summary, SELLF was able to obtain positive rewards while reaching fairness levels similar to an oracle in the selective labels setting.

Model	Lending (Equal. of Opp.)		Criminal Rec. (Acc. Parity)		School admis. (Quali. Parity)	
	Disparity(↓)	Reward(↑)	Disparity(↓)	Reward(↑)	Disparity(↓)	Reward(↑)
PPO	0.38 (± 0.01)	1624.64 (± 14.0)	0.09 (± 0.01)	998.61 (± 2.4)	0.15 (± 0.01)	**1219.68 (± 23.0)**
POCAR	0.37 (± 0.01)	1626.68 (± 12.9)	0.06 (± 0.01)	838.94 (± 7.9)	0.14 (± 0.01)	1203.02 (± 25.4)
POCAR (Oracle)	0.05 (± 0.01)	1156.82 (± 12.4)	0.08 (± 0.02)	920.16 (± 8.2)	**0.12 (± 0.01)**	1069.16 (± 20.5)
FOCOPS	0.41 (± 0.01)	1627.32 (± 14.7)	0.13 (± 0.01)	1000.0 (± 0.0)	0.18 (± 0.02)	1161.7 (± 16.9)
ELBERT	0.30 (± 0.01)	**1630.18 (± 13.2)**	0.12 (± 0.01)	758.07 (± 18.1)	0.18 (± 0.02)	1211.36 (± 24.2)
SELLF (Semi-sto.)	0.05 (± 0.01)	1326.28 (± 16.7)	**0.02 (± 0.01)**	956.07 (± 7.7)	0.14 (± 0.02)	1157.88 (± 24.3)
SELLF (ours)	**0.03 (± 0.01)**	1246.24 (± 14.7)	0.05 (± 0.01)	990.72 (± 3.5)	**0.12 (± 0.01)**	1131.58 (± 26.7)

Table 3: Performance and true disparity averaged over time of agents at the lending (with equality of opportunity) and school admission (with qualification parity) environments. Results are an average of 10 deployment repetitions.

H Additional Results

In this section, we present an analysis of IPW stability during learning and results for different configurations of environments. Summarized results are presented in Tab. 5.

H.1 Analysis of Terms from Theorem 3.3

As discussed in Theorem 3.3 and detailed in Appendix C.4, the error ϵ_t^i of predictor ϕ on the rejected population can be bounded by $\bar{\epsilon}_t^i$, which is composed of two terms:

$$\epsilon_t^i \leq \hat{\epsilon}_{A,\mathrm{w}}^i + 2^{5/4}\sqrt{d_2(D_R^i||D_A^i)}\sqrt[\frac{3}{8}]{\frac{p\log\frac{2N^i e}{p} + \log\frac{4}{\delta}}{N^i}} \tag{73}$$

The first term is the estimate of the error using the accepted population and the weights from IPW. The second term is an increasing function of the Renyi divergence, the pseudo-dimension p of predictor space, the confidence level δ, and is a decreasing function of the number of samples N^i. In our learning algorithm, we aim to reduce this bound by decreasing the error $\hat{\epsilon}_{A,\mathrm{w}}^i$ and the Renyi-divergence $\sqrt{d_2(D_R^i||D_A^i)}$, as we consider that p, δ are fixed, and N^i will increase over learning iterations. In this section, we present an analysis of the values of $\hat{\epsilon}_{A,\mathrm{w}}^i$, $\sqrt{d_2(D_R^i||D_A^i)}$ during learning. To do so, we employ the same setting of the ablation study in Sec. 5, by using the lending environment with equality of opportunity fairness principle, $\beta_1 = 5$, and 5 different values of β_2.

We present average results over 25 random repetitions in Fig. 8, where we display the error measure and Renyi divergence of each group separately. As our results use the bias as a measure of error, that is, $\epsilon(x, i) = \mathbb{E}[\hat{Y} - Y|X = x, Z = i]$ (with no module), a flexible predictor will reach values close to 0. In the figure, we see that for both groups the error is below 0.07 during learning, and gets close to 0 as the value of β_2 increases. Particularly for the privileged group, we see that the bias is higher, indicating an overestimation of $Y = 1$ of the rejected population of the privileged group. When considering the Renyi divergence, our results show that it can present a significantly high value, which could make the bound loose. When the Renyi regularization is not included, that is, $\beta_2 = 0$, the Renyi divergence resulted in the highest value during learning (above 6). These results corroborate the inclusion of the Renyi regularization in our method.

Next, we discuss and evaluate the effect of the complexity of the predictor space on SELLF.

H.2 Predictor with High Pseudo-dimension

The bound presented in Theorem3.3 is also an increasing function of the pseudo-dimension of ϕ, a measure of complexity of the hypothesis space. For linear predictors, as the ones evaluated in Sec. 5, this dimension is equal to $p + 1$, with p being the dimension of the input size. However, for larger neural networks with ReLU activations, the pseudo-dimension will increase with the total number W of parameters and number of layers L by $O(WL\log W)$ (Bartlett et al., 2019). The higher the pseudo-dimension, the higher the number of samples that will be necessary to obtain a tight bound. In this section, we present results where ϕ is a neural network with three linear layers of hidden dimensions $[64, 64, 1]$ and ReLU activation after the first and second layers. We followed the same

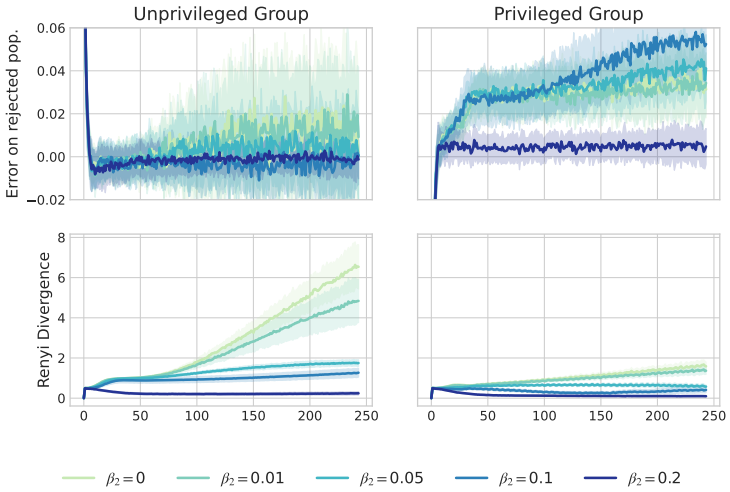
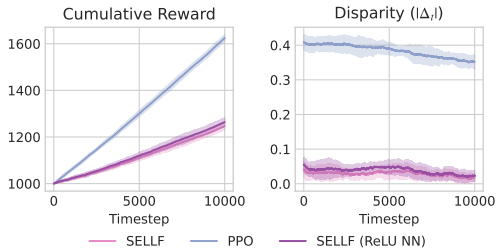
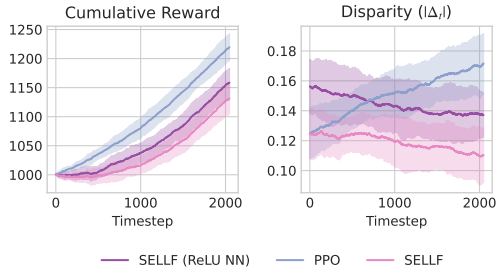


Figure 8: Error term and divergence term of the bound from Theorem 3.3. The Renyi divergence can present high values and make the bound loose.



(a) Lending environment (Equality of Opportunity).



(b) School admission environment (Qualification Parity).

Figure 9: Reward and true disparity over time obtained by optimized agents. Results are obtained with 10 repetitions in the lending environment (a) and the school admission environment (b). Confidence intervals are displayed, but are tight.

experimental settings from Sec. 5, performing hyper-parameter optimization on the values of β_1, β_2. We present a comparison of SELLF with a larger neural network, called SELLF (ReLU NN), with the implementation of SELLF with a linear predictor and the baseline PPO.

Results are displayed at Fig. 9a, 9b and Tab. 4. For the lending environment with the equality of opportunity fairness principle, both SELLF and SELLF (ReLU NN) obtained similar results, satisfying the fairness constraint of 0.05. The results obtained by SELLF (ReLU) on the school admission environment presented higher disparity than the ones obtained by SELLF. Yet, it presented a decrease in disparity over time, surpassing the PPO baseline. As X has 128 features on the school admission environment, ϕ will contain a large number of parameters and a large pseudo-dimension, making it necessary to have a larger number of samples to reduce the bound from Theorem 3.3.

H.3 Learning Stability

We performed a simple ablation experiment to analyze the importance weights $\mathrm{w}(x,i) = D_R^i(x)/D_A^i(x)$ employed by SELLF, as small values of D_A^i can lead to unstable learning. To do so, we evaluated the maximum value $\mathrm{w}(x,i)$ and the minimal value of $P(A[1:K] = 1|x,i) := P(\bigvee_{k=1}^{K} A[k] = 1|X = x, Z = i)$ during learning for different configurations of $\beta_2 \in \{0, 0.01, 0.05, 0.1, 0.2\}$ with fixed $\beta_1 = 5$. We used the lending environment with the accuracy parity fairness principle (the same one used by the ablation study in Sec. 5).

Model	Lending (Equal. of Opp.)		School admis. (Quali. Parity)	
	Disparity(↓)	Reward(↑)	Disparity(↓)	Reward(↑)
PPO	0.38 (± 0.01)	**1624.64 (± 14.0)**	0.15 (± 0.01)	**1219.68 (± 23.0)**
SELLF	**0.03 (± 0.01)**	1246.24 (± 14.7)	**0.12 (± 0.01)**	1131.58 (± 26.7)
SELLF (ReLU NN)	0.04 (± 0.01)	1263.96 (± 17.3)	0.14 (± 0.02)	1158.66 (± 23.9)

Table 4: Performance and true disparity averaged over time of agents at the lending (with equality of opportunity) and school admission (with qualification parity) environments. Results are an average of 10 deployment repetitions.

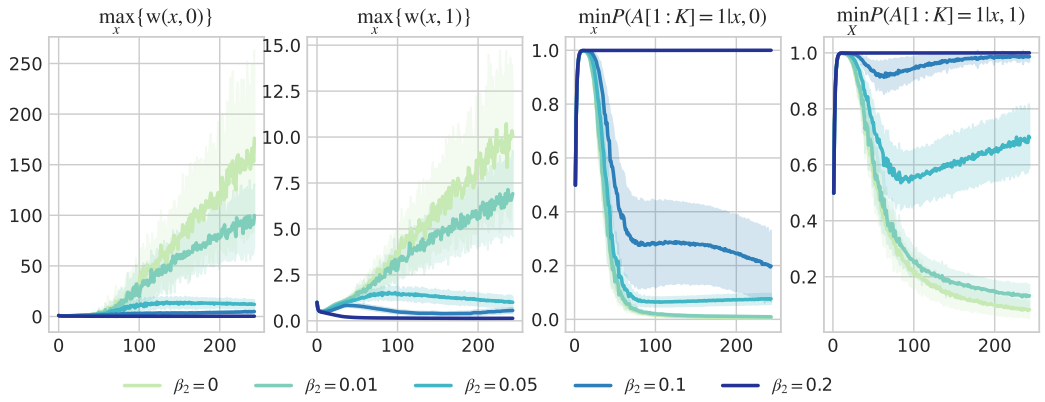


Figure 10: Behavior of importance weights $w(x, z)$ during learning and probability of acceptance at previous iterations with the lending environment with accuracy parity fairness principle.

Fig. 10 presents the results of 25 random repetitions of training, with results displayed separately for each group, where 0 represents the underprivileged group. When $\beta_2 = 0$, the maximum weight of group 0 increases during learning, reaching values higher than 150 at the end. This effect is also present on the group 1, however, reaching values of 10. This difference in weights between groups occurs as they will have different acceptance rates, and the group with the lowest acceptance rate will lead to high values of importance weight. However, as we increase the value of the hyperparameter β_2, the value of $\max_x w(x, i)$ decreases for both groups, reaching very low values when $\beta_2 = 2$. This shows how the Renyi regularization can reduce the maximum value of $\max_w w(x, i)$ and consequently increase learning stability. SELLF calculates $P(A[1 : K] = 1|x, i)$ at each round by sampling 10 policies and calculating the aggregated probability of acceptance by them. With values of $\beta_2 \in \{0, 0.01\}$, this probability gets closer to 0 at the end of learning, as policies are more specialized and tend to only accept a subset of the population. When the weight of the Renyi regularization increases, this effect is reduced. While $\beta_2 = 0.1$, the probability for the unprivileged group reaches 0.2, and with $\beta_2 = 0.2$, it stays fixed at 1 after a few initial iterations.

H.4 Additional Environment: Crime Recidivism

We present results in a crime recidivism environment based on previous works by (Zhang et al., 2020a; Rateike et al., 2024) by leveraging the COMPAS dataset (ProPublica). In this environment, individuals are described by two features, age and prior count, with X having 13-dim. The sensitive attribute is $Z = 0$ if the individual is African American and $Z = 1$ if they are Caucasian. A decision-maker must choose between jail ($A = 0$) or bail ($A = 1$), and is negatively rewarded if they grant bail ($A = 1$) and the individual reoffends ($Y = 0$), and a small positive reward if $Y = 1$. We do so by setting the cost $c = 0.9$. In this example, the only dynamic present is if an individual is granted bail and reoffends, their priors count feature increases by one. In this experiment, we use the accuracy parity fairness principle. We emphasize that this environment is a simplification and does not fully represent real-world dynamics of the criminal justice system.

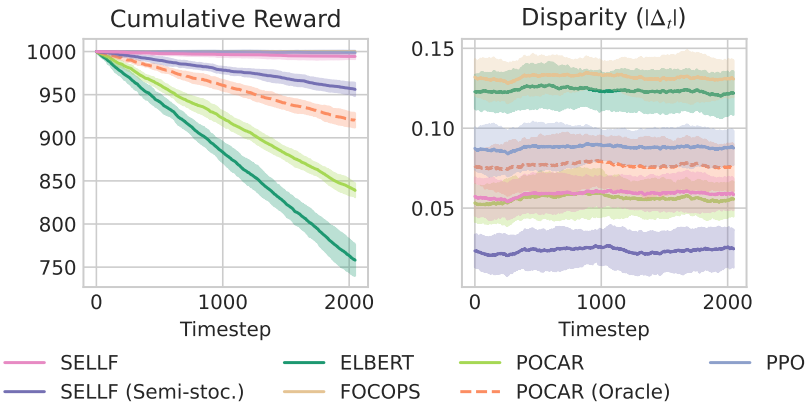


Figure 11: Reward and true disparity (accuracy) obtained in the crime recidivism environment. Results are obtained with 10 repetitions.

Comparative Results We display results in Fig. 11. In this environment, due to the high cost of a false positive (bail a reoffender), no algorithm obtained a final reward above $1,000$. For the same reason, POCAR (Oracle) only obtained results better than PPO when the weight of fairness penalization was increased. Our solution was the only algorithm that resulted in a disparity below 0.05. Interestingly, SELLF (Semi-sto.) obtained the lowest disparity. However, it resulted in a lower reward than standard SELLF.

H.5 Environments with Other Fairness Notions

Lending with Accuracy Parity Fig. 12 presents the results for accuracy parity in the lending environment. In this scenario, SELLF (Semi-sto.) obtained the best results in terms of disparity, followed by FOCOPS. In this setting, PPO was also able to obtain disparity below 0.05. This occurs as the decision-maker's utility and the individual's utility are aligned (both are positively rewarded by setting $A_t = Y_t$). All algorithms, except for POCAR (Oracle) and ELBERT, obtained similar rewards. Particularly in this environment, SELLF presented an increasing disparity over time, starting from 0.05 up to 0.08. This might occur as SELLF increases the acceptance rate of the policy to obtain better confidence bounds on ϕ, leading to accepting individuals with $Y_t = 0$.

Lending with Qualification Parity This environment presents a high initial unfairness of 0.43, and considering the model $\alpha(x, z)$ as presented in Sec. E, accepting individuals with lower scores will lead to decreasing their qualification, as individuals with credit score lower or equal than 2 have more than 50% chance of having $Y_t = 0$. For that reason, no agent was able to present improvements in terms of disparity, including POCAR with oracle access. Interestingly, SELLF obtained a reward higher than PPO in this environment. This might occur due to the incentive for acceptance introduced by the Renyi regularization.

Criminal Recidivism with Equality of Opportunity In this environment, both SELLF (Semi-sto.) and FOCOPS obtained 0 of disparity and 1000 of reward (the same amount as the starting value). This occurs due to the high costs of false positive decisions; both algorithms resulted in an agent that has a 0 acceptance rate. This conservative decision-making might not be ethical in this scenario, as it denies the bail opportunity for every individual. SELLF, POCAR, and POCAR (Oracle) all obtained disparity below 0.05, with similar reward values.

Criminal Recidivism with Qualification Parity Similar to the lending environment with qualification parity, all algorithms reached a similar high-disparity. This occurs as the initial disparity is considerably high, and the decision-maker does not have a significant impact on qualifications. PPO, POCAR (Oracle), FOCOPS, and SELLF obtained similar rewards in this setting.

Model	Lending (Acc. Parity)		Lending (Quali. Parity)	
	Disparity(↓)	Reward(↑)	Disparity(↓)	Reward(↑)
PPO	0.04 (± 0.01)	**1624.64 (± 14.0)**	**0.42 (± 0.01)**	1607.42 (± 15.7)
POCAR	0.06 (± 0.00)	1611.60 (± 15.2)	**0.42 (± 0.01)**	1529.86 (± 13.3)
POCAR (Oracle)	0.08 (± 0.00)	1417.88 (± 21.7)	**0.42 (± 0.01)**	1556.70 (± 13.5)
FOCOPS	0.03 (± 0.01)	1617.72 (± 16.2)	**0.42 (± 0.01)**	1626.5 (± 13.8)
ELBERT	0.53 (± 0.0)	1089.7 (± 6.5)	**0.42 (± 0.01)**	**1627.86 (± 17.8)**
SELLF (Semi-sto.)	**0.02 (± 0.01)**	1627.1 (± 16.1)	**0.42 (± 0.01)**	1484.32 (± 28.8)
SELLF (ours)	0.07 (± 0.01)	1617.54 (± 20.5)	**0.42 (± 0.01)**	1611.66 (± 29.2)
Model	Recidivism (Eq. Opp.)		Recidivism (Quali. Parity)	
	Disparity(↓)	Reward(↑)	Disparity(↓)	Reward(↑)
PPO	0.09 (± 0.01)	998.61 (± 2.4)	**0.13 (± 0.01)**	997.9 (± 2.3)
POCAR	0.03 (± 0.0)	998.72 (± 1.8)	**0.13 (± 0.01)**	907.46 (± 8.9)
POCAR (Oracle)	0.04 (± 0.01)	999.15 (± 1.4)	**0.13 (± 0.01)**	998.8 (± 1.1)
FOCOPS	**0.0 (± 0.0)**	**1000.0 (± 0.0)**	**0.13 (± 0.01)**	**999.91 (± 0.3)**
ELBERT	0.45 (± 0.01)	758.07 (± 18.1)	0.14 (± 0.01)	798.17 (± 15.1)
SELLF (Semi-sto.)	**0.0 (± 0.0)**	999.41 (± 0.9)	**0.13 (± 0.01)**	919.73 (± 10.1)
SELLF	0.04 (± 0.0)	999.29 (± 1.6)	**0.13 (± 0.01)**	997.15 (± 2.7)
Model	School admis. (Eq. Opp.)		School admis. (Acc. Parity)	
	Disparity(↓)	Reward(↑)	Disparity(↓)	Reward(↑)
PPO	0.27 (± 0.02)	**1211.26 (± 13.9)**	0.06 (± 0.01)	1211.26 (± 13.9)
POCAR	0.27 (± 0.02)	1210.78 (± 14.3)	0.07 (± 0.01)	1117.5 (± 15.4)
POCAR (Oracle)	0.05 (± 0.02)	1139.36 (± 13.6)	**0.05 (± 0.01)**	1179.5 (± 15.4)
FOCOPS	0.05 (± 0.01)	1212.0 (± 16.3)	0.09 (± 0.02)	1201.58 (± 14.6)
ELBERT	0.31 (± 0.02)	1226.94 (± 16.8)	0.08 (± 0.01)	**1226.94 (± 16.8)**
SELLF (Semi-sto.)	0.06 (± 0.01)	1189.4 (± 14.1)	**0.03 (± 0.01)**	1205.96 (± 18.6)
SELLF (ours)	**0.04 (± 0.01)**	1161.36 (± 26.3)	0.05 (± 0.01)	1193.28 (± 18.6)

Table 5: Performance of agents at the lending, recidivism, and school admission environments. Results are an average of 10 deployment repetitions.

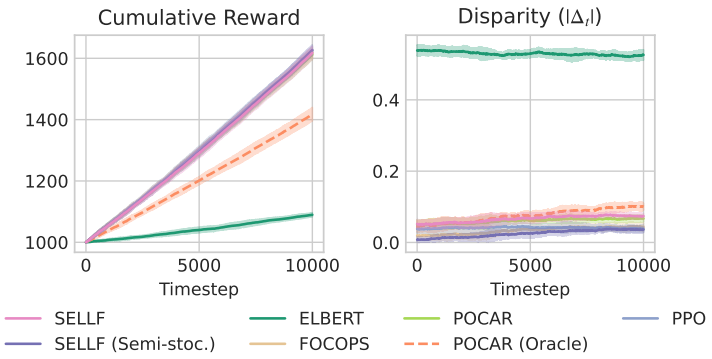


Figure 12: Reward and true disparity (accuracy parity) over time obtained by optimized agents in the lending environment. Results are obtained with 10 repetitions; confidence intervals are displayed, but are tight.

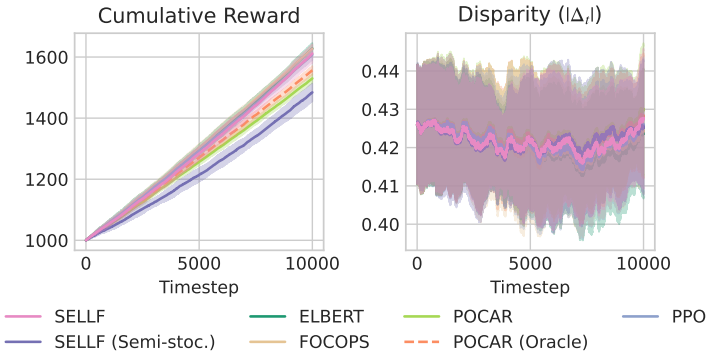


Figure 13: Reward and true disparity (qualification parity) over time obtained by optimized agents in the lending environment. Results are obtained with 10 repetitions.

School Admission with Equality of Opportunity Fig. 16 presents the results for the school admission environment with equality of opportunity. POCAR (Oracle), SELLF and SELLF (Semi-sto.) were able to reach disparity values lower than 0.05, with the lowest value obtained by SELLF. When considering the cumulative reward, SELLF (Semi-sto.) presented slightly higher results than POCAR (Oracle).

School Admission with Accuracy Parity Fig. 17 presents the results for the school admission environment with accuracy parity. In this setting, PPO, ELBERT, FOCOPS, and SELLF (Semi-sto.) reached similar results in terms of cumulative reward, with ELBERT having the highest value. As previously discussed, the accuracy parity notion is a utility measure that behaves similarly to the reward of the decision-maker. For that reason, PPO reached a disparity measure of 0.06. POCAR (Oracle) and SELLF presented similar results, having disparity values equal to 0.05, but were both surpassed by SELLF (Semi-sto.). Despite this variation not having the theoretical guarantees, it consistently presented performance in terms of disparity and reward, similar to the standard variation of SELLF.

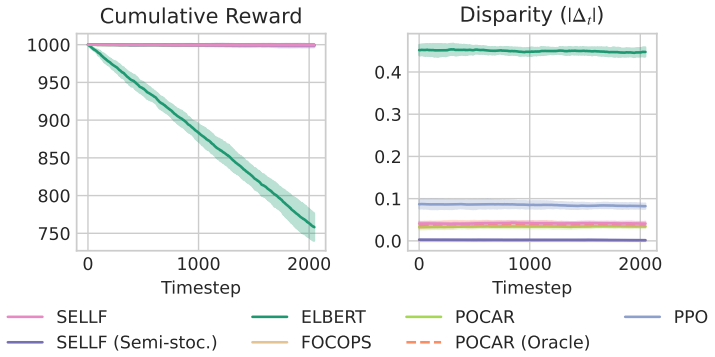


Figure 14: Reward and true disparity (equality of opportunity) over time obtained by optimized agents in the recidivism environment. Results are obtained with 10 repetitions.

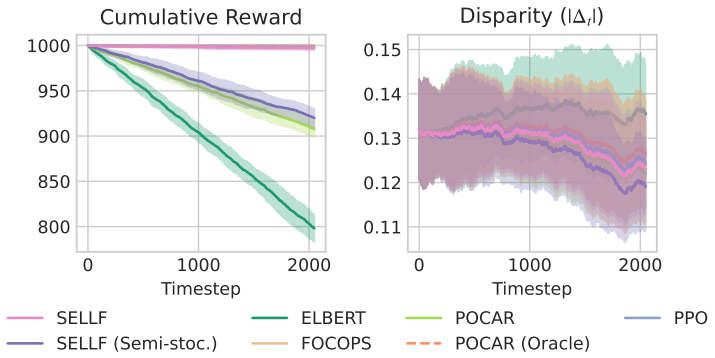


Figure 15: Reward and true disparity (qualification parity) over time obtained by optimized agents in the recidivism environment. Results are obtained with 10 repetitions.

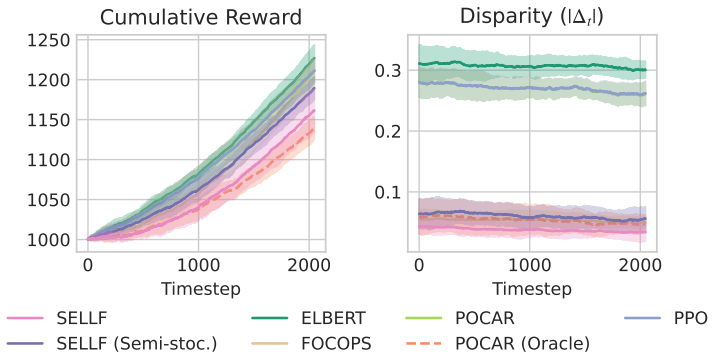


Figure 16: Reward and true disparity (equality of opportunity) over time obtained by optimized agents in the school admission environment. Results are obtained with 10 repetitions.

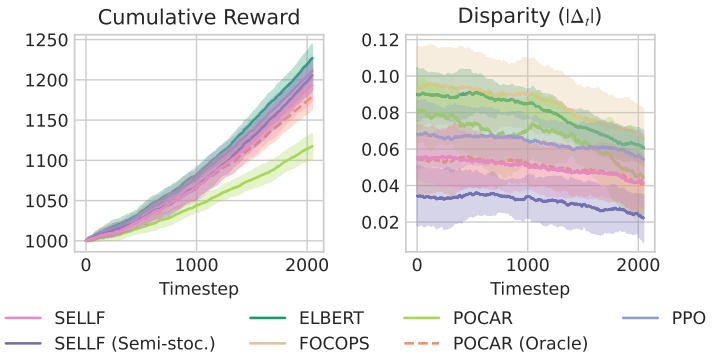


Figure 17: Reward and true disparity (accuracy parity) over time obtained by optimized agents in the school admission environment. Results are obtained with 10 repetitions.